Università degli Studi di Padova

Department of Mathematics "Tullio Levi-Civita"

Master Thesis in Data Science

Micro-estimates of Multidimensional Child Poverty in sub-Saharan Africa

Supervisor
Prof. Lamberto Ballan
Università degli Studi di Padova

Master Candidate
Marina Vicini

Student ID
2021116

Academic Year

2021-2022

To
Sunta and Mirella,
Sandro and Ico,
for your kind hearts
and
everything you taught me

Abstract

Child poverty maps allow governments and other organizations to design policies to track and evaluate their impact in the fight against child poverty. However, reliable data on the geographic distribution of child poverty is scarce, sparse in coverage and expensive to collect. For some countries, the only available measurements are at the country level. In this thesis, we propose to train Machine Learning models to obtain finely grained predictions of child poverty using heterogeneous and publicly available data sources as geographical, demographic and economic georeferenced inputs. Benchmarks of child poverty, computed from nationally representative household survey data, are used as targets to train and calibrate our proposed prediction models. The multidimensional child poverty index has six dimensions: sanitation, water, education, housing, health and nutrition, and is defined such that the predictions can be compared across countries. Using the techniques that are introduced in this thesis, we compute and release a complete and publicly available set of micro-estimates of prevalence, depth and specific poverty dimensions at a 5.2 km^2 resolution for sub-Saharan African countries. Prediction intervals are included to facilitate responsible downstream use. The resulting micro-estimates have the potential of being used to deepen the understanding of the causes of child poverty in sub-Saharan Africa and to gain insights on the impact of future actions.

Contents

Abstract v

List of figures ix

List of tables xi

Listing of acronyms xiii

1 Introduction 1
- 1.1 Child Poverty . . . 2
- 1.2 National Definitions of Child Poverty . . . 3
- 1.3 Internationally comparable child poverty . . . 4
- 1.4 Goal . . . 7

2 Related Works 9
- 2.1 Socioeconomic variable mapping . . . 9
  - 2.1.1 Transfer learning with satellite images . . . 11
  - 2.1.2 Microestimates of relative wealth . . . 13
  - 2.1.3 Refined model . . . 15
- 2.2 Current Child Poverty maps . . . 16

3 Data Collection 21
- 3.1 Ground-Truth Measurements . . . 21
  - 3.1.1 Grid Construction . . . 22
  - 3.1.2 Neighboring approach . . . 24
  - 3.1.3 Missing Values . . . 25
  - 3.1.4 Output . . . 27
- 3.2 Input Data . . . 27
- 3.3 Processing . . . 35
  - 3.3.1 Data scaling . . . 35
  - 3.3.2 Imputing . . . 36

4 Models 39
- 4.1 Decision Trees . . . 42
- 4.2 Ensemble Methods . . . 43
- 4.3 Gradient Boosting . . . 44

4.4 XGBoost 48
4.5 LightGBM 49

5 Experiments 51
5.1 Metrics 52
5.2 Spatial Cross Validation 53
5.3 Feature Importance 53
5.4 Prediction Intervals 55

6 Results 57
6.1 Model dependent on country 57
6.2 Generalizable model 58

7 Conclusion 77

References 81

Acknowledgments 93

Listing of figures

2.1 Relation between asset-based poverty and nightlight intensity 12
2.2 Cross-validated R^2 result for consumption expenditure and asset-based poverty 13
2.3 Pipeline of the refining model. 16
2.4 Performance of the refining model for a subset of countries. 17
2.5 Results of the refining model in Sierra Leone. 17
2.6 Distribution of number of deprivations with MODA methodology 18
2.7 Country level prevalence and depth through MODA methodology 19

3.1 Locations of hexagon with DHS measurements in Nigeria 22
3.2 Countries with recent DHS surveys 23
3.3 H3 hexagonal grid 23
3.4 Example of the neighboring approach 25
3.5 Missing values per dimension 27
3.6 Correlation between child poverty dimensions 28
3.7 Distribution of the output variable 37
3.8 Correlation between features and response variable 38

4.1 Variance bias trade-off 41
4.2 Example of decision tree 43
4.3 Boosting and bagging 44
4.4 Boosting algorithm 45
4.5 Regression losses 48
4.6 XGBoost level-wise growth 49
4.7 LightGBM leaf-wise growth 50

5.1 Methodology pipeline diagram 51
5.2 Spatial cross validation in Nigeria 54

6.1 Comparison of XGBoost and LightGBM performance 63
6.2 Boxplot of R^2 distribution for each dimension 64
6.3 Performance of prevalence and depth per country 65
6.4 Focus of predictions of prevalence and depth in Nigeria 66
6.5 Performance of model trained and evaluated in the same country 67
6.6 Countries in test set 68
6.7 Spatial and standard cross validation 69
6.8 Visualizations of prevalence in sub-Saharan Africa 70

6.9 Visualizations of prevalence's prediction intervals in sub-Saharan Africa . . . 71
6.10 Visualizations of depth in sub-Saharan Africa 72
6.11 Visualizations of depth's prediction intervals in sub-Saharan Africa 73
6.12 SHAP values for generalizable model 74
6.13 State level 75
6.14 Caption 75

Listing of tables

1.1 Indicators for child poverty dimensions 8

3.1 Countries without recent DHS surveys 24
3.2 Countries with recent DHS surveys 26
3.3 Input data 35

6.1 R^2 scores for all countries 59
6.2 R^2 scores for all countries for six dimensions 60
6.3 R^2 results of Generalizable model 61

Listing of acronyms

AdaBoost	Adaptive boosting
CISI	Critical Infrastructure Spatial Index
CNN	Convolutional Neural Network
DHS	Demographic and Health Survey
DSSG	Data Science for Social Good
EFB	Exclusive Feature Bundling
FVU	Fraction of the Variance Unexplained
GBDT	Gradient Boosting Decision Tree
GDP	Gross Domestic Product
GEE	Google Earth Engine
GOSS	Gradient-Based One-Side Sampling
GPS	Global Positioning System
HRSL	High-Resolution Settlement Layer
IQR	Interquartile Range
IWI	International Wealth Index
KNN	*k*-Nearest Neighbors
LMIC	Low- and middle- income country
MICS	Multiple Indicator Cluster Surveys
MAPIE	Model Agnostic Prediction Interval Estimator
MODA	Multidimensional Overlapping Deprivation Analysis
MSE	Mean Square Error

NDVI	Normalized Difference Vegetation Index
NDWI	Normalized Difference Water Index
OSM	OpenStreetMap
PCA	Principal component analysis
PDSI	Palmer Drought Severity Index
SHAP	Shapley additive explanations
STC	Save the Children
UNICEF	United Nations Children's Fund
VIIRS DNB	Visible Infrared Imaging Radiometer Suite Day/Night Band
WKT	Well Known Text
XGBoost	eXtreme Gradient Boosting

1 Introduction

In the world 1.2 billion children suffer from child poverty [1]. Child poverty has long lasting effects on children, both on the short term and on the long term. Lack of realization of basic needs causes a child to not enjoy his/her rights and not reach his/her full potential, causing threats to the child's health and well being. Moreover, adults with a history of childhood abuse and/or neglect tend to have lower levels of education, lower wages and less employment [2]. The effects of child poverty are even more long lasting, such that it will take 4 to 5 generations (i.e. around 150 years) for a family that grows up in poverty to reach the average national level of income [3].

Given its alarming consequences, poverty is the first point of the "Sustainable Development Goals", adopted in 2015 by the United Nations member states. The goal is to "end poverty in all its forms everywhere". This objective is subdivided in more specific targets, we can focus on the second one that claims: "By 2030, reduce at least by half the proportion of men, women and children of all ages living in poverty in all its dimensions according to national definitions" [4]. In the target above, we want to highlight several points that will be discussed more in depth along the chapter. There is a specification of children as separate from adults, since child poverty is different from general poverty, and poverty is specified in all its dimensions given its multidimensionality. Moreover, there is the need for national definitions in terms of poverty, and measurements of the proportion of poverty in terms of sex and age.

However, not all countries have developed their definition of child poverty. Moreover, these measurements are not comparable since they have been computed based on different definitions. Thus, a new definition of child poverty has been proposed by UNICEF with the goal of being internationally comparable [5]. One of the objectives of this definition of child poverty is policy design by organizations such as UNICEF and private organizations such as Save The Children.

In this introductory chapter, we will start by defining child poverty in Section 1.1, then we will talk about how to give a national definition of poverty in 1.2, that will be useful to understand the internationally comparable definition of child poverty in Section 1.3. At the end of the chapter we will be able to specify the goal of this thesis in Section 1.4.

1.1 Child Poverty

Let us start by defining child poverty [5]:

> Child poverty is the lack of public and private material resources to realize their rights constitutive of poverty.

Rights of poverty are those that directly depend on material resources for their continued realization. For example, sanitation and housing are rights constitutive of poverty, while privacy and happiness are not.

We need to measure child poverty separately from adult poverty because the needs are different, starting with education. Children should not work to earn a living, they depend on adults for support, care. Moreover, children are between 25% to 50% of the national population, and in some countries, even more than half. Hence, not having a measure of poverty specific to the child would lead to a miscalculation of national poverty, leading to incorrect policy design and incorrect assessment of policy impact [6].

Child poverty should be independent of monetary income. It is neither a cause or consequence of monetary income. Child poverty needs to be measured because it affects the children and "the deprivation in these rights is what makes the child poor" [5]. Moreover, measuring child poverty in terms of monetary income is not only inadequate [7], but it could be harmful. For example, if the household income is the result of child labor, then lowering the percentage of households in monetary poverty is damaging for the children. Or if income increases because adults work overtime, that could lead to children being neglected and ending up in unsafe situations since their parents are never home.

1.2 National Definitions of Child Poverty

To reach the Sustainable Development Goals, each country should have their national definition of poverty. To tackle this point, United Nations Development Programme, UNICEF and World Bank have jointly produced a guide for countries to measure multidimensional poverty (poverty measured in terms of rights and not solely based on income) [?].

Such as for child poverty, also general poverty should not be measured based on income, since monetary poverty does not provide the full picture. With multidimensional poverty, it is important to assess the deprivation of each dimension, and then to understand how these deprivations overlap between each other.

The government should be the entity to measure the process of child poverty, because its participation as an official entity grants legitimacy to the process, facilitating the use by other stakeholders. The definition and measurements of poverty should be rigorous, transparent, institutionalized, sustainable, and useful.

The guide specifies the methodology to measure multidimensional child poverty, and it is based on identification and aggregation of the deprivations, that is:

1. Identify a set of relevant dimensions of poverty and define indicators for each dimension.

2. Determine criteria to assess deprivations.

3. Define a satisfactory threshold for each indicator.

4. Classify each individual (or household) as deprived or not in each dimension based on the criteria and the thresholds.

5. Sum the number of deprivations with the respective weights of each indicator.

6. Define a poverty cut-off.

7. Aggregate the results of individuals (or households).

This 7-step guideline can be used identify multidimensional poverty measures for each country, and similar steps have been taken to determine a definition of internationally comparable child poverty.

1.3 Internationally comparable child poverty

While each country is supposed to have its own way to measure child poverty, this is not always the case. Moreover, even with these definitions we cannot use them to compare poverty in different countries. Therefore, we want to introduce another study carried out by UNICEF to address this problem [5]. These estimates of child poverty are comparable because they use the same dimensions, the same indicators and the same threshold. The focus of these measurements is for supra-national poverty assessment and they are not indicated for national ones.

For the internationally comparable estimates of child poverty, four principles have been followed:

1. estimates at the individual child level
2. the dimensions considered must be rights constitutive of poverty
3. all dimensions are equally weighted
4. measure how poor children are

In the next subsections, we can dive more into details about each point.

Individual Child Level

The first principle says that the estimates should not be a disaggregation of a household measure, but measured directly at the individual level. Doing the opposite -measuring poverty at the household level, and then disaggregating- could not be beneficial to the child, if the household is not considered below the poverty line because of reasons just regarding adults with no effects on the children's lives [6]. For example, if an adult is not unemployed anymore, and does not take the child to doctor's visits.

Rights constitutive of poverty

Human rights are rights that allow one to take part in ordinary living activities customary to the society in which they belong; these have been referred to as capabilities [8]. While rights constitutive of poverty are those whose realization is blocked by lack of material resources. The rights constitutive of poverty approach remarks that child poverty estimates center on individual children, since they are, differently from households, rights-holders.

Other measurements such psycho-social and emotional deprivations, neglect, violence are not material, and therefore not considered in the context of child poverty, but they are part of the Wellbeing and Quality of Life of Children. Moreover, data limitations is another important factor in the selection of indicators and dimensions, since in some household surveys there are not the full set of indicators, or not all indicators are asked to all children. Hence, the definition includes 6 dimensions that are [9]:

> education, health, housing, nutrition, sanitation, and water.

For each dimension included, we need to specify the indicators and the thresholds, made explicit in Table 1.1, that have been selected following these criteria:

- Simplicity: consider only one indicator per dimension to avoid imbalance across dimensions.
- Maximize country coverage: using indicators that are available in all data surveys.
- Validity: using indicators based on material deprivations and not emotional or spiritual ones.
- Reliability: accurate measurements.
- Internationally agreed criteria.
- Feasibility to separate severe and moderate deprivation,

Another key point to take into account is that in absence of knowledge, no imputation needs to be made. This data limitation leads to underestimating poverty, that is, however, better than overestimating it.

As for the national definitions of child poverty, there is a two-step approach: identification (which children are deprived in which dimension), and aggregation (individual summary measure of the child's information).

Dimensions equally weighted

Human rights are the birthright of each human being, and therefore are universal [10]. Declared internationally in 1948 by the Universal Declaration of Human Rights in 1948, all rights are of equal validity and importance [11]. While equal importance of rights holds for every human being, the guideline of Committee on the Rights of the Child focuses more specifically about children's rights saying that "they are indivisible and interrelated, and that equal importance should be attached to each and every right recognized therein" [12].

Hence, from equal importance of all human rights we can derive equal weighting of dimensions, that means that all dimensions have the same weights. These are the reasons for it [13]:

- Not only weighting implies a ranking between rights, but it explicitly says how much one right is more important than the other.
- Having weights could cause a person with more deprivations to be less poor than one with just one deprivation (if this deprivation has a greater weight). Even if weights are decided based on experts' opinions or a focus group with the people in question (considered "poor"), they would still be arbitrary.
- A more statistical approach would be following the principle of indifference or insufficient reasoning [14]. The principle says that in absence of any sufficient evidence, the agent should have the same degree of belief towards all outcomes.
- Literature advises strongly against weighting of indexes, showing that applying weights does not provide any gain in information [15]. Moreover, since the dimensions are all rights constitutive of poverty and therefore are all based on lack of material resources, the indexes are expected to be correlated between each other. This points to saying that it is not worth weighting [16].
- Another reason is ease of communication: lack of weighting makes the definition of multidimensional poverty easier to interpret, making it more transparent.
- For the Capabilities approach, there is no trade off between rights.
- Lastly, applying Ockham's razor we follow the more parsimonious option of no weighting.

We can lastly note that equal weighting holds only for dimensions, and not for indicators. However, in these definitions there is only one indicator per dimension, so there are no concerns for weighting in the different indicators.

Measuring Depth

The last point focuses on, not only measuring the prevalence of child poverty, but also the depth. As a child poverty cut-off, the number of deprivations to consider a child poor, in this approach, is 1. Thus, a child is multidimensional poor if he or she is deprived in at least one dimension. We need to focus also on the number of deprivations a child is deprived in, since prevalence does not paint the whole picture, and it would underestimate improvements in children's lives (if a policy reduces the number of deprivations from 5 to 2, the prevalence does not change, but the situation of the child is improved. Other important measurements should be the number of children that suffer from one deprivation, two deprivations, three ect. The whole distribution of deprivations is required to comprehend poverty at the individual level.

It is important to address equity and to explore the differences between boys and girls, geographic locations, parent's formal education and other disparities.

1.4 Goal

Having high-resoluted measurements of child poverty is an important step for policy-making and to track how poverty changes through time. Traditionally a geographic distribution of poverty is computed combining a household survey with a broader survey, such as a census. However, this method produces official statistics [17], and household surveys are expensive and slow to collect, and not available for all countries. Moreover, the discussion for the lack of data and the research for alternative ways to estimate poverty is accentuated regarding child poverty, since the subset of survey data available to measure ground truth data is almost halved, since in many countries children are about 50% of the population.

The objective of this thesis is to compute finely grained measurements of internationally comparable child poverty in all the 48 countries in Sub-Saharan Africa, including the countries that lack DHS surveys in recent years. The estimates have a resolution on average of 5.16 km^2, and they will follow an hexagonal grid. Measurements will be available only in areas with at least 30 children for privacy reasons. The variables that have been estimated will be: the average number of dimensions deprived, the proportion of children deprived in at least 2 or 3 dimensions, the proportion of children deprived in sanitation, water, housing and education. These variables will be predicted by applying Machine Learning models to georeferenced data such as satellite images, economic features, population bands ect.

In chapter 2, previous works in the field are going to be analyzed, in chapter 3 we focus on the data collection part of the pipeline, in 4 we dive into the theoretical concepts of modeling, and in chapter 5 we focus on different aspects of the experiments, whose results will be shown in chapter 7.

Dimension	Unit of Analysis	Severe Deprivation Definition	Moderate Deprivation Definition (includes severe deprivation)
Shelter	Children under 17 years of age	Children living in a dwelling with five or more persons per sleeping room.	Children living in a dwelling with three or more persons per sleeping room.
Sanitation	Children under 17 years of age	Children with no access to a toilet facility of any kind (i.e. open defecation, or pit latrines without slabs, hanging latrines, or bucket latrines, etc).	Children using improved facilities but shared with other households.
Water	Children under 17 years of age	Children with no access to water facilities of any kind (i.e. using surface water or unimproved facilities such as non-piped supplies).	Children using improved water sources but more than 15 minutes away (30 minutes roundtrip)
Nutrition	Children under 5 years of age	Stunting (3 standard deviations below the international reference population).	Stunting (2 standard deviations below the international reference population).
Education	Children between 5-14 years of age	Children who have never been to school.	Children who are not currently attending school.
	Children between 15-17 years of age	Children who have not completed primary school.	Children who are not currently attending secondary school (or did not complete secondary school).
Health	Children 12-35 months old	Children who did not receive immunization against measles nor any dose of DPT.	Children who received less than 4 vaccines (out of measles and three rounds of DPT).
	Children 36-59 months old	Children with severe cough and fever who received no treatment of any kind.	Children with severe cough and fever who did not receive professional medical treatment.
	Children 15-17 years old	Unmet contraceptive needs.	Unmet contraceptive needs (using only traditional methods)

Table 1.1: Dimensions of child poverty with respective indicators, and thresholds for moderate and severe deprivations. Table from [5].

2
Related Works

Measuring poverty has been the focus of research for over a century, since the first survey of living standards was taken more than 100 years ago in England [18]. Since then, many studies have been conducted on how to define poverty and how to map it. On this last point, we can observe that conventional methods to predict poverty in small areas require a population census and some other nationally representative survey [19]. These techniques have been criticized in terms of coverage and accuracy [20]. Since then, new approaches to predict poverty have been implemented [21, 22, 23], applying newer and more sophisticated methods from statistical models to Machine Learning and Deep Learning.

Literature for poverty predictions is very rich, however, it is more focused on general poverty rather than child poverty. Hence, we focus on methods on how to produce finely grained maps for other socioeconomic variables with Machine Learning in Section 2.1, while in Section 2.2 we show current results for child poverty maps in sub-Saharan Africa.

2.1 Socioeconomic variable mapping

While literature for high resolution child poverty maps is scarce, we can analyze the research on new techniques and methodology to produce how resolution maps for wealth, poverty based on income or consumption, well-being, community happiness. These studies vary for the grid construction, methods used and data sources. On this last point, we can differentiate the studies in feature-based models and image-based models.

For the first category data sources can be very different. Mobile phone data are a rich source of information, implemented in a program targeting study in Afghanistan to focus on ultra-poor households [24]. Another way to include mobile phone information was taken by Khan and Blumenstock [25], modelling the mobile phone network as a multi-view graph in a semi-supervised learning scenario to predict poverty. At the individual level, poverty and wealth have been predicted in Rwanda from the past history of mobile phone use. After combinatorially engineering factors of communications, the data has been fed to an elastic net model [26]. Mobile phone information have been used also in combination with other data sources, as we can observe in a study in Bangladesh where a hierarchical Bayesian geostatistical model was implemented using as input mobile phone metadata, such as basic phone usage, top-up patterns and social networks, and environmental and physical metrics, such as vegetation indexes, nighttime lights, climate, distance to roads [27]. An interesting point of the study is the grid construction: predictions were provided on a Voronoi tessellation grid, based on mobile cell tower positions, providing more localized information in urban areas and sparser information in rural areas. The effects of geographical factors have been analyzed with some spatial regression techniques in Kenya [28]. While the study is localized to the country, we can see that other research teams made use of geographical factors in their studies [27, 23, 29, 30]. For example, geographical features such as nighttime light intensity, Normalized Difference Vegetation Index (NDVI), land surface temperature, built-up areas and point of interests have been passed to a Random Forest model to predict poverty [30]. Another interesting data source is textual data, such as tweets or Wikipedia articles. The former have been used to define a finely grained map of "gross community happiness" at the community level in London [31]. While the latter have been processed to predict economic development at the community level, extracting textual information from geolocated Wikipedia articles [32].

We can focus on the second group of models, that are the image based one. A first approach is based on applying transfer learning to train daytime satellite images to predict nighttime light intensity with a convolutional neural network (CNN). From this model, features have been extracted and fed to Machine Learning models to predict poverty [21, 22, 33]. The contribution of Jean *et al.* will be explained more in depth in Section 2.1.1. While other studies use nighttime light intensities not as the output of the CNN model from which features are extracted, but training two separate models, joined at the end in a final fully connected layer [34], or in two separate models whose output are fed into each other to obtain better training data [29]. Other studies exploits satellite images, without considering nighttime light intensity. The minimum well-being poverty line and the well-being poverty line were estimated in

Mexico applying a end-to-end CNN to Planet imagery and Digital Globe imagery [35]. They use GoogleNet as architecture, fine-tuning the weights from ImageNet. In Belize, alternative satellite sources, such as MODIS and Landset, have been used since they are freely available and released as open-source [36]. All the bands, not just the visible ones, are being used in the study. The model used is an ensemble model of Ridge regression, elastic net regression, random forest, extreme gradient boosted trees. Reducing even more the data sources needed, a study uses as predictors just the the NDVI, that measures vegetation greeness: reflecting on the importance of including agricultural production as predictors [37]. The images are from NASA's Terra satellite. While within these studies we have the most performing models, the models are not interpretable. The scope of these maps is policy making, so it is important for the policy makers to trust and understand the process.

Most studies predict poverty just for one country, with the exception of a couple of studies [34, 32, 22, 23, 29]. Now we want to focus on the contribution of three important studies: Jean *et al.* predict poverty from satellite images, Chi *et al.* use satellite images and other data source and Lee *et al.* propose a new model to refine training data obtaining state of the art results.

2.1.1 Transfer learning with satellite images

Starting from [21], Jean *et al.* introduce transfer learning methods from satellite imagery to estimate poverty [22]. This study is focused on 5 African countries: Malawi, Nigeria, Rwanda, Tanzania, and Uganda. The measures of poverty predicted are two: one based on assets, and another based on consumption expenditures. The first measure has been computed from DHS surveys, as the principal component of survey responses of multiple questions about assets ownership [38]. The second one has been obtained from the World Bank's Living Standards Measurement Study surveys [39]. The authors use daytime satellite images and nighttime light intensities as input. Nightlight intensity values go from 0 to 63, and they have been binned in three classes 0-2, 3-34, 35-63. These values have been taken from the United States Air Force Defense Meteorological Satellite Program satellites, that have been processed by National Oceanic and Atmospheric Administration's National Geophysical Data Center to estimate global human-generated lighting between 20:30 and 22:00 local time every day [40]. The resolution of this dataset is 1 km. In later studies, an updated dataset is used for this purposes, that is the Visible Infrared Imaging Radiometer Suite Day/Night Band (VIIRS DNB) [33, 34, 23, 29]. Since the location provided by data surveys used is jitter to ensure privacy, they extracted night-lights estimates within a 10 km × 10 km square centered in the provided coordinates, to which the mean value was assigned. The daylight images have been collected from the Google Static Maps at zoom level 16, with a resolution 2.5 m per pixel. Hence, to match the nighttime luminosity resolution, the images are 400 × 400 pixels. For each household cluster, 100 images were extracted to convert a 10 km × 10 km area, while for larger states 25 evenly spaced images were extracted per cluster.

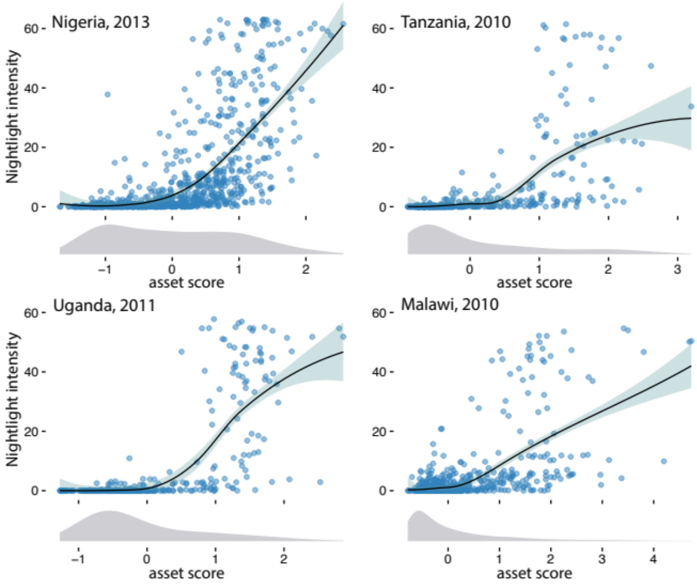


Figure 2.1: Relation between asset-based poverty and nightlight intensity at cluster level for Nigeria, Tanzania, Uganda and Malawi. Image taken from [22].

Now, we can focus on the modelling part. The method follows two steps, first through transfer learning training a model to predict a proxy for poverty, and then extract features from the model to actually predict poverty. The proxy in the first part is used since poverty maps are scarce, and the variable is nighttime light intensity, noisy but easily obtainable proxy, reasoning its correlation to urban developments. The relations between the two variables can be observed in Figure 2.1. The authors start by fine-tuning an 8-layer CNN model (VGG F) pretrained on ImageNet, that consists of a model trained to classify images in 1000 different categories. While predicting the categories, the model learns low-level image features, such as edges and corners. These features can be used in other tasks, even not related to the categories learnt in the initial scope of the model. Hence, the CNN model is fine-tuned to predict nighttime light intensities, starting from daytime satellite imagery. Given the resolution of the data, for each cluster the CNN model is run 100 times, obtaining 100 features vectors. The average of this vector acts as the input features to predict poverty.

This CNN is used as a feature extractor, summarizing the high dimensional input into a low dimensional set of features predictive of variation of nighttime lights. These features are used to train a ridge regression model to estimate wealth, applying regularization to avoid overfitting. The model explains from 55% to 75% of the variation in average household asset wealth, and for 41% to 56% of consumption expenditures across the 5 countries considered as can be seen in the diagonal of Figure 2.2. While in the other entries of the image, we can see how the train model on one country would perform on the others, to assess how the model would perform in other countries without survey data.

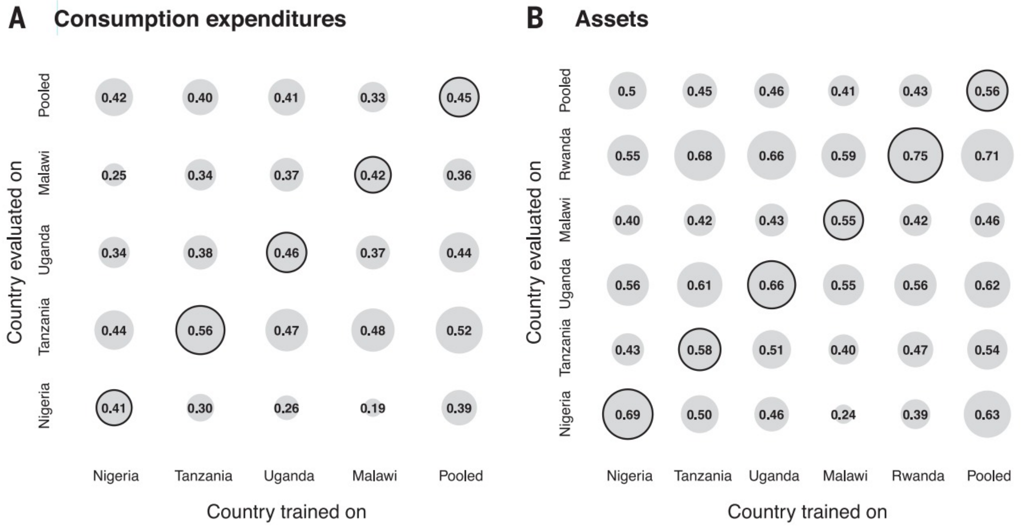


Figure 2.2: Cross-validated R^2 of a model trained on one country and evaluated on another for consumption expenditure (A) and asset-based poverty (B). The countries considered are Nigeria, Tanzania, Uganda, Malawi, and then also an additional column with pooled results from the four countries. Image taken from [22].

Starting from this work, another study has obtained similar results using publicly available, freely distributively satellite images from Landsat 7, at a lower resolution [33].

2.1.2 Microestimates of relative wealth

Chi *et al.* computed available micro-estimates of relative wealth for all low- and middle- income countries, located in Africa, Asia, Europe and Latin America [23]. These predictions are freely available for 91 countries. The authors used the Microsoft Bing tile system [41], and the grid has a resolution of (2.4×2.4) km^2. The output to predict is relative household wealth, computed from DHS survey data, taking the first principal component of 15 questions regarding assets and housing characteristics. The measure is relative to the country, and within each country it has mean 0 and standard deviation equal to 1. The model uses as input several data sources: both quantifiable geospatial variables and features extracted from images. The first set of features are road density, land cover, elevation, slope, precipitation, population, connectivity variables such as cell towers, WiFi access points, mobile devices. Connectivity, nighttime radiance and population density are found to be some of the most important features. We can observe that the connectivity data are proprietary to Meta and not publicly available. While for image-based features, the authors followed an approach similar to [22]. The images are 256×256, and have a resolution of 0.58 m per pixel. The images are downsampled averaging each 16×16 block obtaining a resolution of 9.375 m per pixel. Then, the images are fed to a pre-trained CNN. The network used is ResNet-50, that has been trained on 3.5 billion Instagram images to predict correspective hashtags [42], without tuning the parameters In the second to last layer, they extract a 2048 feature vector, then compress it with PCA, obtaining a feature vector with 100 dimensions (the first 100 components explained 97% of the variance). Since DHS surveys jitter the GPS location, to ensure that the input data cover the true location, for each location, they compute a population-weighted mean of the input data of a 2×2 or 4×4 grid of cells centered on the DHS location. The different grid size depends on urban or rural areas, since the location is jittered differently for urban and rural areas. The model used for the prediction is a gradient-boosted regression tree, tuning the hyperparameter with three different cross-validation techniques: K-fold cross validation, leave-one-country-out cross-validation and spatially stratified cross-validation. The performance of the model is on average 71% R^2 for cross-country estimation and 77% R^2 for single country estimations. To validate the results the authors use census data and four independent data sources specific to certain countries. An interesting point of this paper is the error modeling. To estimate model performance, they fit a linear regression model on the absolute value of the model's residual (on the cells that have ground truth information), using as predictors some observable characteristics of it, such as input features of the predictive model (except the ones based on images), features regarding the availability of ground-truth data nearby and country-level characteristics. Predictions and error estimates are available for cells with at least 30 people for privacy reasons.

2.1.3 Refined model

A more recent work produces high resolution poverty maps for 25 sub-Saharan African countries, comparing single country and cross-country estimations [29]. The resolution is 1 square mile ($1.6 \times 1.6\,\text{km}^2$). Here the grid is built differently, and it is focused on villages, identified as populated places combining information through OpenStreetMap and United Nations Office for the Coordination of Humanitarian Affairs. The poverty measure used is based on assets. Starting from DHS data, they compute an International Wealth Index (IWI) [43], that allows for cross-country comparison for a fixed time frame. This index is highly correlated with the original DHS wealth index. The methodology employs two sets of models: a feature-based one and an image-based one. The data sources considered are: OpenStreetMap (OSM), VIIRS DNB nighttime lights dataset, the High-Resolution Settlement Layer (HRSL) and daytime satellite images [44, 45, 46, 47]. From OSM, they extract length of road and distance to closest road or junction, number of junctions and building and total building area. From VIIRS DNB dataset, that has 15 arcsec resolution, they compute for each cell six summary statistics: max, mean and median of luminosity, ratio of zero luminosity and upper and lower third luminosity. From the HRSL dataset, they extract population estimates, compensated with WorldPop data for countries for which it is not available.

After an overview of the data sources, we can dive more deeply into the model, which can be seen in Figure 2.3. The first step is training an XGBoost (eXtreme Gradient Boosting) model using OSM data, nighttime luminosity and population density as input and the IWI as output. After fine-tuning the hyperparameters with Bayesian optimization, the better model is chosen between the cross-country and single country estimator. Here the trained model is applied to all the populated places in the 25 countries considered, obtaining an estimate of wealth. These values are then fed to the third step, that is a customized CNN. These types of models need abundant data, a strategy previously adopted is nightlight as a proxy as in [22, 21], but nighttime luminosity is limited, especially in rural areas. The CNN model returns a probability distribution of the 1 square mile cell to be rich, upper-middle class, lower-middle class or poor. Classification has been chosen over regression to capture multi-faceted qualitative geospatial features that may go lost otherwise. The model is iterated over time and we can see how the results improve after each iteration in Figure 2.4. The performance is state-of-the-art for poverty maps, achieving on average an R^2 of 86% for cross-country estimation and 88% for single country estimation. The results for a single country are shown in Figure 2.5.

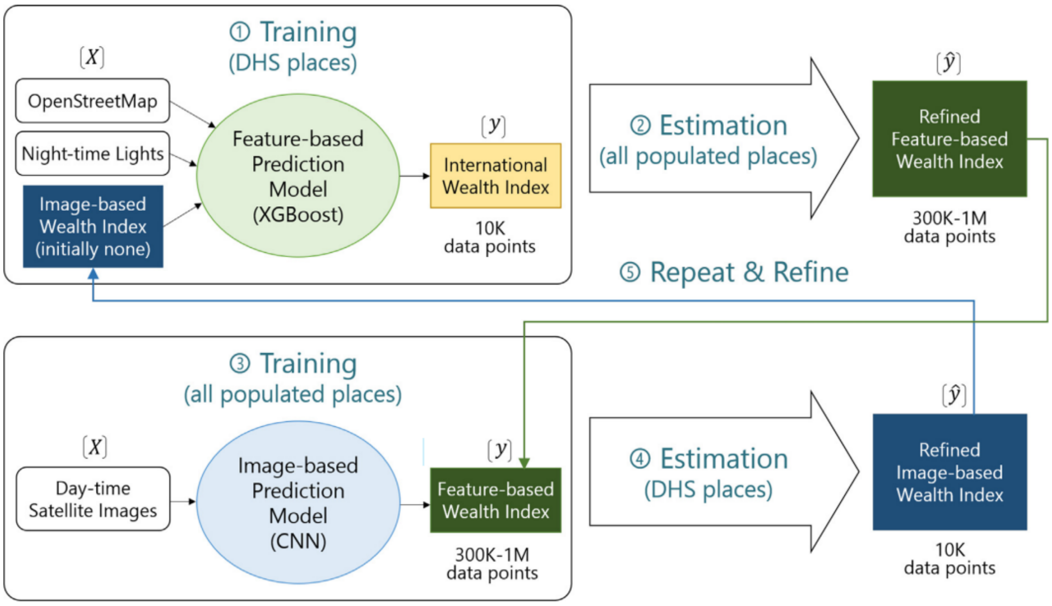

① Training
(DHS places)
[X]
OpenStreetMap
Night-time Lights
Image-based Wealth Index (initially none)
Feature-based Prediction Model (XGBoost)
[y]
International Wealth Index
10K data points
② Estimation
(all populated places)
[ŷ]
Refined Feature-based Wealth Index
300K-1M data points
⑤ Repeat & Refine
③ Training
(all populated places)
[X]
Day-time Satellite Images
Image-based Prediction Model (CNN)
[y]
Feature-based Wealth Index
300K-1M data points
④ Estimation
(DHS places)
[ŷ]
Refined Image-based Wealth Index
10K data points

Figure 2.3: Pipeline of the refining model of [29]. At step (1) an initial XGBoost model is trained on OSM and nightlight data to predict IWI. After training, the model is used to predict IWI for all populated places at step (2). At (3), a CNN model is trained on day-time satellite images to classify a discretization of the IWI output. After training, the model predicts the label for the data used in training at step (1), and these values are fed to a new XGBoost model and the process is repeated. The two models improve each other at each iteration providing more refined training data. Image taken from [29].

2.2 Current Child Poverty maps

In this section we want to focus on the current results of child poverty mapping. An interesting work has been done through the Multidimensional Overlapping Deprivation Analysis (MODA) methodology [48]. This approach has another measurement of child poverty that is still child-centered and multidimensional. The dimensions included are water, sanitation, housing, nutrition, health, education and information. The first three are measured irrespective of age, while housing and nutrition is measured for children below 5, and education and information only for children above 5. Within this framework a child is considered multidimensional poor if he/she has at least two deprivations over five.

The authors use data from DHS and MICS surveys for their analysis, that are separated for age groups, and rural/urban. The analyses are available at the country level and at the sub-Saharan level. The dimensions included are similar, and we can look at the results to have a country level overview of child poverty. The deprivation distribution of multidimensional child poverty can be seen in Figure 2.6. The plots are separated for children below five and over

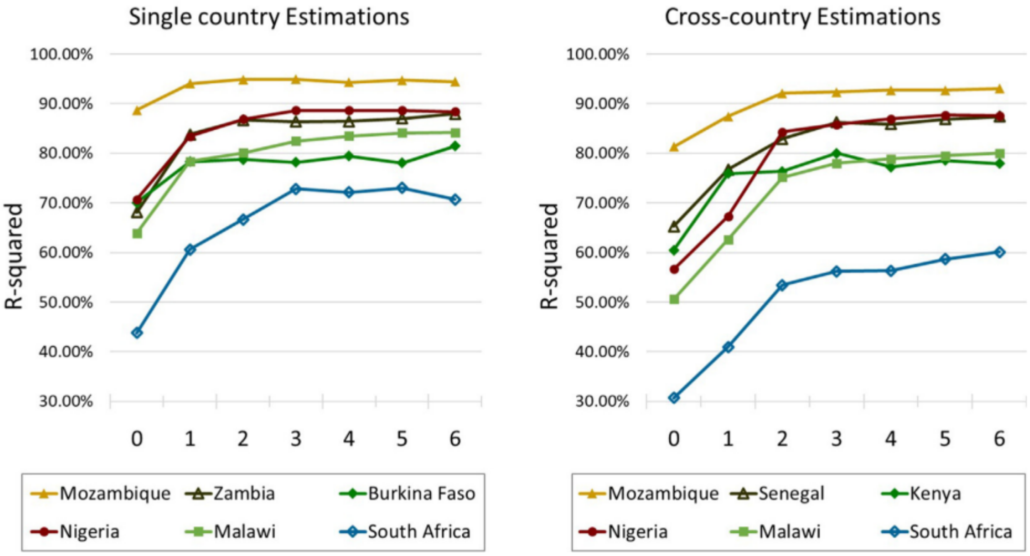


Figure 2.4: Improvement of R-square through each estimation of the refining process in 6 countries: Mozambique, Zambia, Burkina Faso, Nigeria, Malawi and South Africa. Image taken from [29].

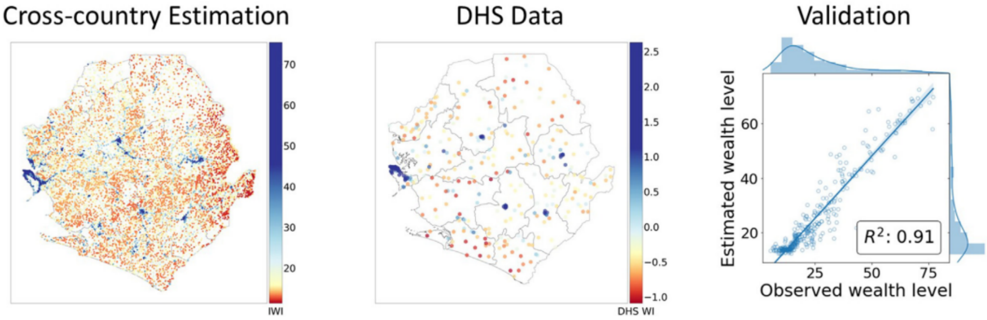
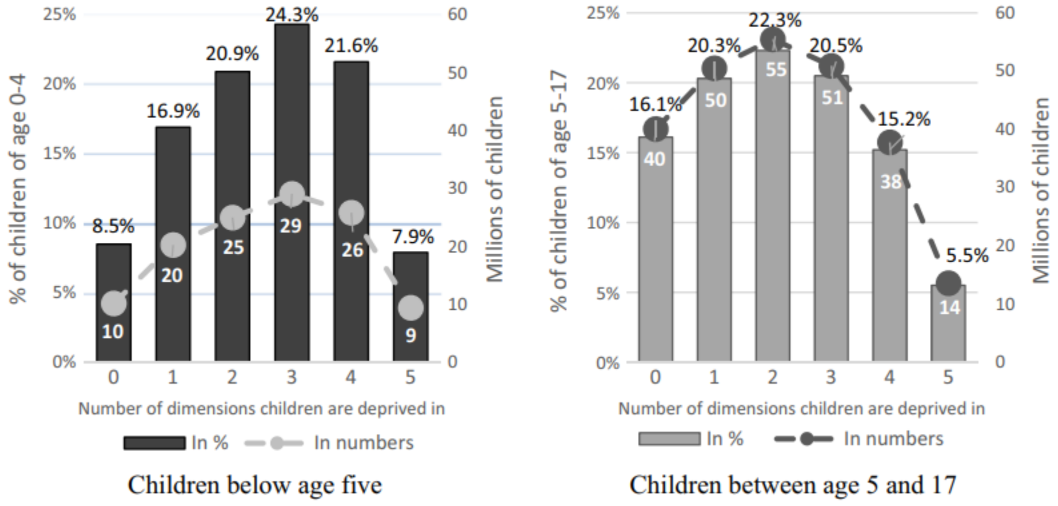


Figure 2.5: Validation of estimations of wealth in Sierra Leone. From the left, cross-country predictions for 13,040 populated places, 336 DHS cluster poverty output, validation plot, achieving a 91.12% R-squared.Image taken from [29].

five, and in total 86.4% of all children experience at least one deprivation, and 67% of children are multidimensionally poor. There are 30 countries considered, and this leads to 247 millions of children.

Then we can analyze the results for singular countries, observing large variation in prevalence in Figure 2.7, as 30% of children are multidimensionally poor in Gabon and 90% in Ethiopia. Focusing on depth, the average number of deprivations among children with at least one deprivation varies from 1.7 in Gabon and Eswatini to 3.4 in Chad and Ethiopia. Overall, we can observe that deprivation intensity and prevalence are generally positively correlated.



Children below age five

Children between age 5 and 17

Figure 2.6: Distribution of the number of deprivations children suffer from, by age-group with MODA methodology. Image taken from [48].

For the remaining countries 14 countries in sub-Saharan Africa without DHS or MICS data, the authors fitted a ordinary least square regression model using the Human Development Index, urban population and population size as predictions, estimating that 64.4% children in the 44 countries of Sub-Saharan Africa are multidimensionally poor, that corresponds to 291 millions of children.

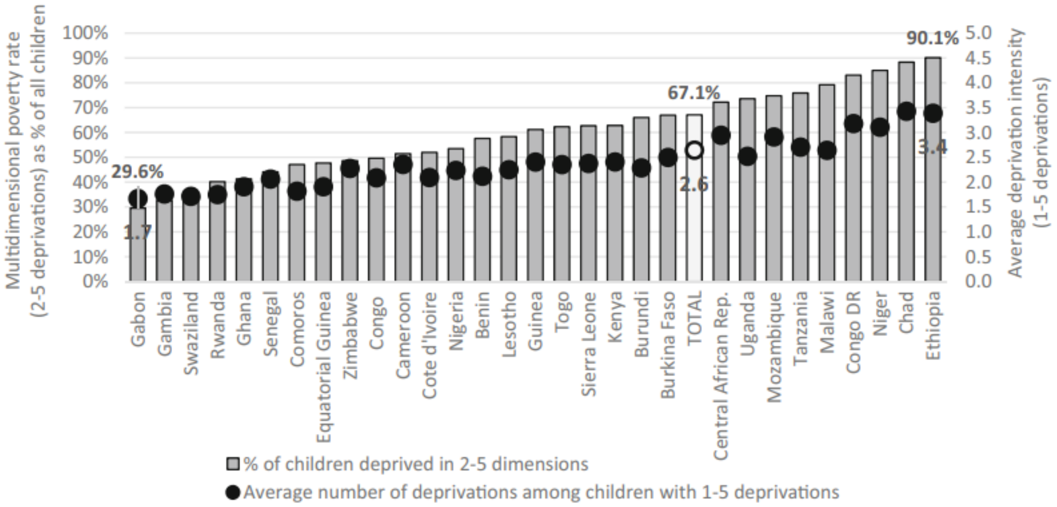


Figure 2.7: Multidimensional poverty prevalence and depth computed for 30 countries in sub-Saharan Africa with the MODA methodology. Image taken from [48].

3 Data Collection

Collecting data is an intensive and fundamental step in a data science project, and we divide the following section in two parts: output and input. In section 3.1, we will analyze how the ground-truth measurements of child poverty are computed and how the grid is built, while in section 3.2, we will analyze which data sources are used as input and how they are aggregated. Lastly, in Section 3.3

3.1 Ground-Truth Measurements

Measurements of child poverty multidimensional poverty were computed from traditional face-to-face surveys with 299,977 unique households living in 14,443 villages in 25 different low- and middle- income countries (LMICs). These Demographic and Health Surveys (DHSs), which are independently funded by the US Agency for International Development, contain detailed questions about the economic circumstances of each household and make it possible to compute a standardized indicator of the average asset-based wealth of each village [49, 38]. From the 1980s, 300 surveys have been carried out in just under 100 countries. Weights are assigned to each household to make the aggregation representative to national level. The survey provides a wide range of indicators to monitor population, health and nutrition. We can see an example of the spatial distribution of the people surveyed in Nigeria in Figure 3.1.

The DHS surveys have been processed by Save The Children to compute the indexes of the child poverty dimensions according to the indicators indicated in Table 1.1. The most

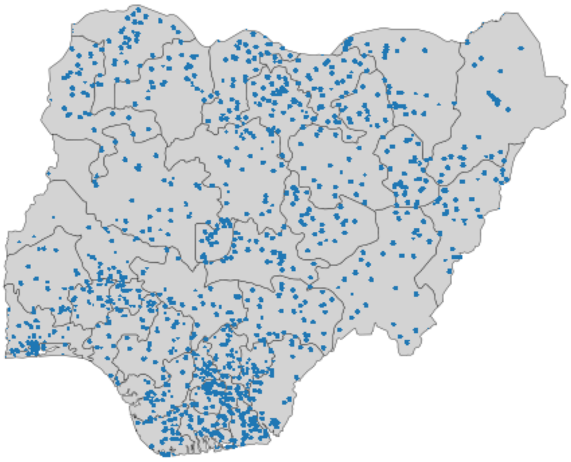


Figure 3.1: Locations of hexagons with DHS measurements in Nigeria. Each hexagon has on average an area of 5.16 km^2.

recent surveys analyzed can be found in Figure 3.2. For each dimension, a binary value indicates whether the child is deprived or not in that dimension.

Another survey that could be used in the future is the Multiple Indicator Cluster Surveys (MICS), which is an international household survey developed and funded by UNICEF to monitor poverty indicators for children. MICS data do not make public the GPS locations, so these surveys can be used for a country level analysis. With STC preprocessing, the countries with available MICS data are: Central African Republic (2019), Chad (2019), The Democratic Republic of the Congo (2018), Congo (2015), Côte d'Ivoire (2016), Eswatini (2014), Gambia (2018), Ghana (2017), Guinea-Bissau (2019), Lesotho (2018), Madagascar (2018), Mauritania (2015), Sao Tome and Principe (2019), Togo (2017).

3.1.1 Grid Construction

To compute a finely grained index of child poverty, we partition the Earth into identifiable grid cells, using Uber's grid system of hexagons [50]. As we have seen in the literature, there are different strategies for computing a grid, that can be a uniform or not uniform grid. An example of uniform gridding is the Voronoi grid or a grid based on postal codes [27]. However, the cells vary in terms of size and area, and they may be subject to change. While for not-uniform, we observe in literature many grids with square cells. Even if this polygonal shape is easy to deal with images, it does not have the best properties for grid construction. For example, squares have two types of neighboring cells: one where they share one vertex and one where they share

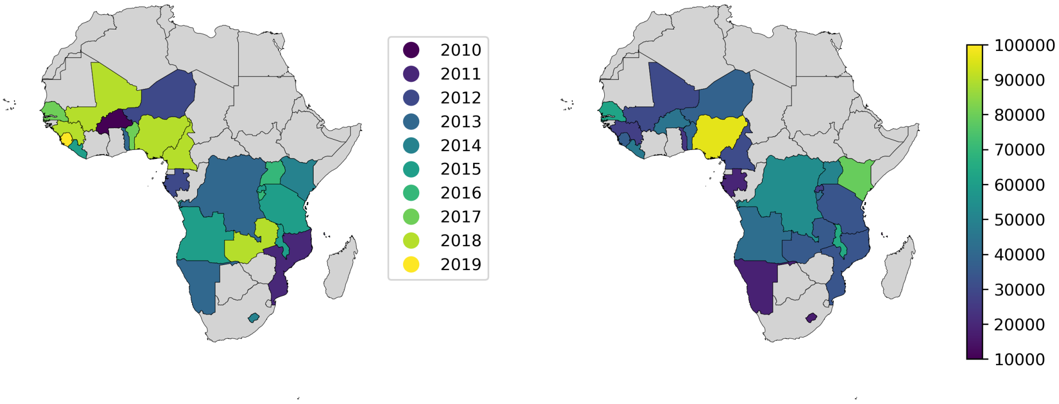


Figure 3.2: In the plot of the left, there are mapped the countries with DHS surveys and the year it has been carried out. On the right there is the number of children surveys per country.

one edge. On the contrary, hexagons have only one type of neighbors, providing an advantage for clustering and dealing with neighbors. The hexagonal shape is chosen due to uniformity of neighbors and reduce sampling bias from edge effects, which is attributed to a high perimeter-area ratio.

Uber provides 16 resolutions for the grid, and we use the seventh, where each hexagon has on average an area of 5.16 km^2. This resolution allows predictions to be finely grained to be used by policy-makers, while respecting the privacy of the households.

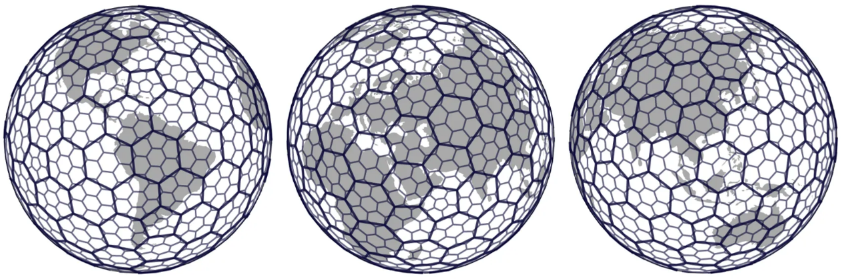


Figure 3.3: Illustrative example of H3 hexagonal grid. Image taken from [50].

The DHS data include GPS coordinates and so we map each individual to the respective hexagonal cell, and the value for the cell is computed taking the mean. Hence, the output represents the proportion of children deprived in that dimension, and it is a value between 0 and 1. While for depth, the target variable represents the average number of deprivations a child on average has per hexagon. For each dimension, we set a threshold on the number of surveys necessary to include the hexagon in the training set to 30. The amount of observations per hexagon changes based on the dimension, since each dimension has a different amount of missing values.

In Table 3.1, we observe the number of hexagons per country of the countries with no DHS survey.

Country	Code	# hexagons
Botswana	BWA	97,688
Cabo Verde	CPV	899
Central African Republic	CAF	121,928
Chad	TCD	224,373
Congo	COG	64,775
Côte d'Ivoire	CIV	83,903
Equatorial Guinea	GNQ	5,737
Eritrea	ERI	23,454
Eswatini	SWZ	3,029
Ethiopia	ETH	232,333
Gambia	GMB	2,266
Ghana	GHA	59,764
Guinea-Bissau	GNB	7,387
Madagascar	MDG	112,606
Mauritania	MRT	202,481
Mauritius	MUS	407
Sao Tome and Principe	STP	274
Seychelles	SYC	189
Somalia	SOM	102,263
South Africa	ZAF	220,528
Sudan	SDN	328,724
South Sudan	SSD	120,393
Zimbabwe	ZWE	63,573
tot		2,078,974

Table 3.1: Countries without recent DHS surveys and the number of hexagons in that country.

3.1.2 Neighboring approach

70% of the individuals from the DHS survey are located in rural areas and 30% in urban ones. To ensure the privacy of the subject of the DHS survey, DHS jitters the GPS location. Urban clusters are displaced up to 2 km, and rural clusters are displaced up to 5 km [38]. A randomly selected 1% of rural clusters are displaced up to 10 km.

Hence, the exact GPS location of the household may not fall in the hexagon it has been assigned to. To take this fact into account, data from each household has been copied to the neighboring hexagons in urban areas and in two neighboring hexagons in rural one. This method ensures that the true location of the household surveyed is covered. In areas where more surveys fall, the mean has taken. This procedure is supported by an assumption of smoothness among neighboring cells. A visual representation of this method can be seen in Figure 3.4.

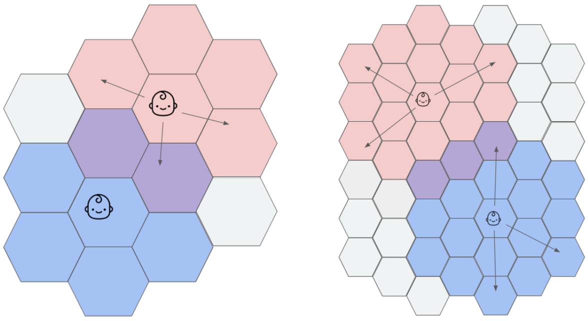


Figure 3.4: Example of the neighboring approach in urban areas to the left and in rural areas to the right. The neighboring approach consists in replicating the output to the k-ring neighboring cells as a way to augment the data and account for GPS jitter in DHS surveys. For rural areas k is equal to 2, and for urban areas k is equal to 1.

In Table 3.2, we can see the number of hexagons per country and the number of hexagons with ground-truth output with and without the neighboring approach, for the countries that have DHS surveys.

3.1.3 Missing Values

In the poverty indicators computed by STC from the DHS surveys, there are some missing values. Some values are missing because those dimensions are not relevant to the child. For example, the dimension education is relevant only for children older than 5, while nutrition only for children younger than 5 and health for children of range 1-5 and 15-17.

However, there are indexes where data is missing and the dimension is relevant to the child. The lack of response to certain DHS variables may be caused because the question was not asked (due to the interviewer error) or because the respondent did not want to answer. DHS highlights the importance of not making up answers, and STC and UNICEF explicitly says that it is better to underestimate poverty than to overestimate it. So no imputation techniques were

Country	Country Code	# hex w output	# hex w output (neigh)	# hexagons
Angola	AGO	519	5,983	215,093
Benin	BEN	468	4,864	26,302
Burkina Faso	BFA	498	7,070	63,999
Burundi	BDI	489	3,668	4,258
Cameroon	CMR	364	4,127	96,805
Comoros	COM	116	361	379
DR Congo	COD	468	6,742	386,149
Gabon	GAB	220	2,276	52,659
Guinea	GIN	362	5,083	56,744
Kenya	KEN	1,181	13,926	102,898
Lesotho	LSO	314	3,088	5,550
Liberia	LBR	561	6,489	25,331
Malawi	MWI	761	7,921	15,369
Mali	MLI	304	4,566	268,851
Mozambique	MOZ	521	6,996	131,056
Namibia	NAM	261	3,272	146,574
Niger	NER	396	5,891	219,041
Nigeria	NGA	1,296	16,678	190,861
Rwanda	RWA	440	3,272	3,821
Senegal	SEN	515	5,941	40,385
Sierra Leone	SLE	464	5,820	17,545
Tanzania	TZA	546	7,568	149,873
Togo	TGO	274	3,473	13,544
Uganda	UGA	634	8,554	37,757
Zambia	ZMB	488	6,583	120,545
tot		12,460	150,233	4,470,363

Table 3.2: Countries with recent DHS surveys, the respective number of children surveyed per country, the number of hexagons with the response output with and without the neighboring approach.

used on the output values. The number of hexagons available for each dimension is shown in Figure 3.5. We can observe that there are no missing values regarding sanitation, housing and water, while this is not the case for education, health and nutrition. While the lack of data for education is not that significant, it is a concern for health and nutrition where there are respectively 779 and 810 hexagons with ground truth data without introducing the neighboring approach in whole sub-Saharan Africa.

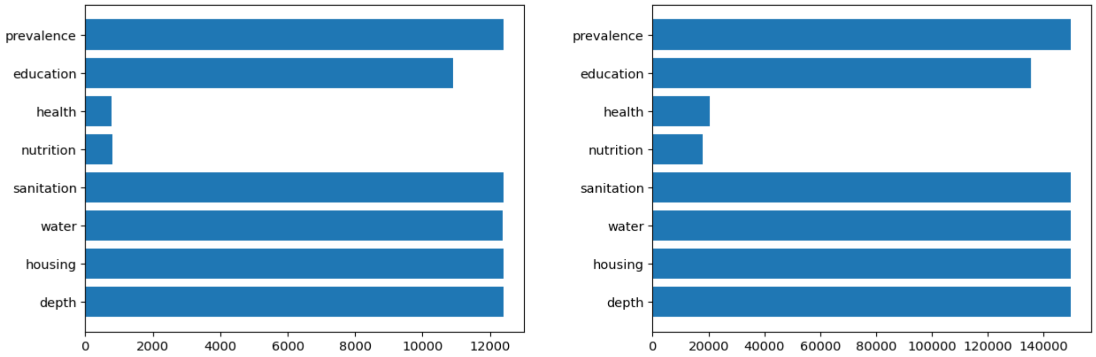


Figure 3.5: Number of available hexagons for each dimensions at the sub-Saharan level, without the neighboring approach (to the left) and with it (to the right). We can notice that nutrition and health have the most missing values, since they are relevant only for a small age group. The amount of data for the proportion of children deprived in at least 1, 2, 3 or 4 dimensions is the same as the number of data of depth.

3.1.4 Output

The output variables are: depth, prevalence, proportion of children with at least 2, 3, 4 deprivations, proportion of children with at least three deprivations, proportion of children deprived in sanitation, water, housing and education. In Figure 3.6, we can see how the different dimensions are correlated between each other. All dimensions are positively correlated between each other. As expected, depth is highly correlated with prevalence and the proportion of children with 2, 3, 4 deprivations that are still correlated with each other. Housing, nutrition and health are the least correlated with the other variables and between each other.

Moreover, we can observe the need to measure all these different dimensions. In Figure 3.7, we can observe the distribution of each output. Prevalence is skewed towards 1, while health, nutrition, housing, water, education and depth are skewed towards 0. For sanitation we find a U-shaped distribution, where the most frequent values are 0 and 1.

3.2 Input Data

The input data are georeferenced data to represent physical, economical, demographic characteristics of each hexagon. A comprehensive list of all the variables can be seen in Table 3.3. Now we will analyze in depth each data source.

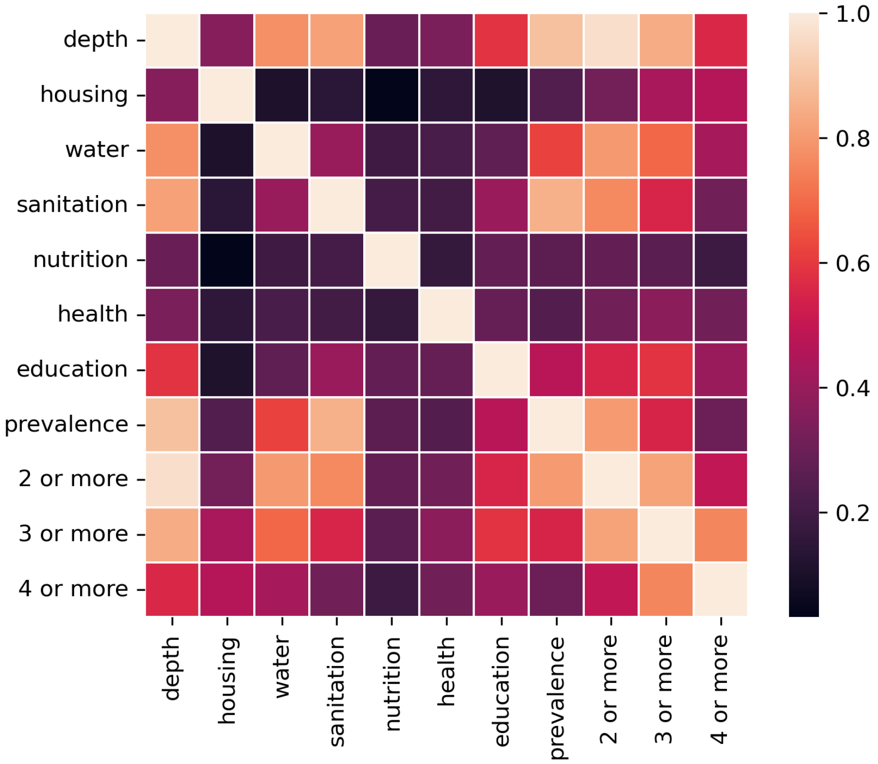


Figure 3.6: Correlation between child poverty dimensions. We can observe that all variables are positive correlated among themselves. Nutrition and health are the least correlated to the other dimensions, and prevalence and depth among the most.

Conflict Zones

The Uppsala Conflict Data Program (UCDP) is a program at the Uppsala Universitet in Sweden, that collects data on organized violence [51, 52]. The dataset used is the most disaggregated one, that covers individual events of organized violence, that can be defined as "phenomena of lethal violence occurring at a given time and place". The disaggregation geolocalize the events at the village level. The UCPD datasets on organized violence are updated yearly.

Open Street Map

Open Street Map is an open source geographical database, built through crowdsourced volunteered geographic information [44]. OpenStreetMap data are used by thousands of websites and mobile applications [53], plus many academic studies [54]. Through the Overpass API, used to extract OSM data, we sum the road length of each hexagon, dividing it for its area, obtaining road density.

Ookland Open Data

Global fixed broadband and mobile (cellular) network performance map tiles, allocated to zoom level 16 web mercator tiles (approximately 610.8 m by 610.8 m meters at the equator). Data is provided in both Shapefile format as well as Apache Parquet with geometries represented in Well Known Text (WKT) projected in EPSG:4326. Download speed, upload speed, and latency are collected via the Speedtest by Ookla applications for Android and iOS and averaged for each tile. Measurements are filtered to results containing GPS-quality location accuracy.

OpenCellID

Open Database of Cell Towers is collected by Unwired Lab, a small company that works in the geolocation space [55]. OpenCellID is a collaborative community that collects GPS positions of cell towers, computing the average of the position of the received radio signal of a GSM base station. From the database, we extract information about GSM, UMTS and LTE, that are respectively 2G, 3G and 4G mobile network technologies,

Critical Infrastructure

Critical infrastructure is of extreme importance for the functioning of society and for socio-economic development. A study aggregates OpenStreetMap data and measure its global spatial intensity through the Critical Infrastructure Spatial Index (CISI) [56]. The available resolution is 0.10×0.10 degrees.

Economics

Economical variables such as Gross Domestic Product (GDP) have been computed at high-resolution by a team at Aalto University in Finland [57]. The dataset contains global measures GDP per capita for the 25-year period of 1990–2015. We included the latest available measures.

Wealth

The relative wealth index is an estimation of wealth of one micro-region with respect to other areas in the same country. The index has been computed for low and middle income countries, creating a grid of (2.4×2.4) km^2. This work has been assessed by Meta's Data for Good team in collaboration with the University of Berkeley [23].

Commuting Zones

The Data for Good team at META has identified geographic areas where people live and work, called commuting areas [58]. These areas are different from traditional boundaries of cities, counties or states and can help identify how people move and interact, giving insights on local economies. From the data we obtain geometries of the commuting zone and the area, the road length and population of each commuting zone.

The following data sources are extracted through Google Earth Engine.

Topography

The Shuttle Radar Topography Mission provides an elevation dataset, from which slope can be computed [59]. The elevation is measured in meters.

Vegetation and Water

From Landsat 8 of the U.S. Geological Survey, normalized difference water index (NDWI) and normalized difference vegetation (NDVI) index are included. The first has a composite value with a time span of a year, while the second one over the time span of 32 days. Both indexes have a resolution of 30 m. The use of NDVI is found in literature as a predictor of poverty [37].

Precipitation

TerraClimate is a dataset computed by the Climatology Lab of the University of California Merced, that contains climatic water balance for global terrestrial surfaces [60]. The dataset has a monthly temporal correlation and a 4 km spatial resolution. From the dataset, evapotranspiration (measured in millimeters), precipitation accumulation (measured in millimeters) and Palmer Drought Severity Index (PDSI) are collected.

Global Precipitation Measurement is an international satellite project that records and estimates data observations of rain and snow globally [61]. With a resolution of 11 km, we extracted the "merged satellite-gauge precipitation estimate", measured in millimeters per hour.

Human Settlement

Global Human Settlement Layers project supported by the Joint Research Centre, the European Commission and Directorate-General for Regional and Urban Policy, measures the multitemporal build-up presence with 38 m resolution to describe the human presence in the planet [45, 46, 47].

Healthcare

A collaboration between MAP (University of Oxford), Telethon Kids Institute (Perth, Australia), Google, and the University of Twente, Netherlands have mapped accessibility to healthcare, computing the travel time to the nearest hospital or clinic, with or without motorized transport, at a spatial resolution of 927.67 m [62].

Nightlight

Colorado School of Mines computed monthly average radiance composite images, using nighttime data from the Visible Infrared Imaging Radiometer Suite Day/Night Band. With a spatial resolution of 463.83 m, we included average DNB radiance values and cloud-free coverages, that is the the total number of observations that went into each pixel, since for certain areas there are no quality data, due to cloud cover or solar illumination. Nighttime intensity is an important variable in literature for the prediction of poverty [21, 22].

Pollution

MODIS Terra and Aqua and Multi-angle Implementation of Atmospheric Correction (MAIAC) produced a data product that measures the land aerosol optical depth, with a spatial resolution of 1 km a temporal resolution of one day [63]. From this collection, we extracted blue band (0.47 μm) and green band (0.55 μm) aerosol optical depth over land. These can be used as measures of pollution.

Population

World Pop estimates population dataset per each country, releasing a new dataset every year [64]. It also provides estimations for age/sex structure per each country at a 100 m spatial resolution. World Pop measures population in a constrained and unconstrained manner. The first one is using a mask of the settlement layer and the second one does not use any mask, since the population layer would carry the errors of the settlement mask. Hence, we use the unconstrained demographic maps [65].

We used population estimation for 2020. The dataset is available in GeoTIFF format with resolution of 3 arc seconds (that corresponds to approximately 100m at the equator), and the projection is Geographic Coordinate System, WGS84. The files include age information in the following age groups: from 0 to 12 months, from 1 to 4 years, and then every 5 year age-group, until 80+ group.

Name	Description	Source
`hex_code`	hex code	
`country_code`	country code	
`geometry`	polygon geometry wkt	
`n_conflicts`	Number of conflicts	Conflict Zones
`length_km`	Road Length [km]	Open Street Map
`area_km2`	Area hexagon [km^2]	Open Street Map
`road_density`	Density of roads in hexagon [1/km]	Open Street Map
`avg_d_kbps`	Average download speed [kilobits per second]	Internet Connectivity
`avg_u_kbps`	Average upload speed [kilobits per second]	Internet Connectivity
`GSM`	Mobile tower for 2G network	Mobile Cell Tower
`UMTS`	Mobile tower for 3G network	Mobile Cell Tower
`LTE`	Mobile tower for 4G network	Mobile Cell Tower
`avg_signal`	Average Signal	Mobile Cell Tower
`africa`	Critical Infrastructure Spatial Index	Critical Infrastructure
`ec2019`	Electricity consumption of 2019	Electricity
`GDP_PPP_1990`	Gross Development Product for 1990	Economics
`GDP_PPP_2000`	Gross Development Product for 2000	Economics
`GDP_PPP_2015`	Gross Development Product for 2015	Economics
`2019gdp`	Gross Development Product for 2019	Economics
`rwi`	Relative Wealth index	Wealth
`rwi_error`	Error of Relative wealth index	Wealth
`name_commuting`	Name of the commuting zone	Commuting Zones

win_population_commuting	Population of commuting zone	Commuting Zones
win_roads_km_commuting	Total length of roads of commuting zones	Commuting Zones
area_commuting	Area of commuting zone	Commuting Zones
elevation	Elevation	Topography
slope	Slope	Topography
evapotrans	Evapotranspiration	Precipitation
precipiacc	Accumalation of precipitation	Precipitation
precimean	Average precipitation	Precipitation
precistd	Standard deviation of precipitation	Precipitation
pdsi	Palmer Drought Severity Index	Precipitation
ndvi	Normalized Difference Vegetation Index	Vegetation
ndwi	Normalized Difference Water Index	Water
water_surface	Water surface	Human Settlement
no_built	Land no built-up in any epoch	Human Settlement
build_2000_2014	Built-up from 2000 to 2014 epochs	Human Settlement
build_1990_2000	Built-up from 1990 to 2000 epochs	Human Settlement
build_1975_1990	Built-up from 1975 to 1990 epochs	Human Settlement
build_prior_1975	Built-up up to 1975 epoch	Human Settlement
cnfd	Confidence of the settlment class	Human Settlement
accessibility	Travel time to the nearest hospital	Access to healthcare
accessibility_walking_only	Travel time to hospital using non-motorized transport	Access to healthcare
avg_rad	Average DNB radiance values.	Nightlight
cf_cvg	Number of cloud-free observations	Nightlight
Optical_Depth_047	Blue band (0.47 µm) aerosol optical depth over land	Pollution
Optical_Depth_055	Green band (0.55 µm) aerosol optical depth over land	Pollution
population	Population	Population
M_0	Male population between 0 and 12 months	Population

M_1	Male population between 1 and 4 y.o.	Population
M_5	Male population between 5 and 9 y.o.	Population
M_10	Male population between 10 and 14 y.o.	Population
M_15	Male population between 15 and 19 y.o.	Population
M_20	Male population between 20 and 24 y.o.	Population
M_25	Male population between 25 and 29 y.o.	Population
M_30	Male population between 30 and 34 y.o.	Population
M_35	Male population between 35 and 39 y.o.	Population
M_40	Male population between 40 and 44 y.o.	Population
M_45	Male population between 45 and 49 y.o.	Population
M_50	Male population between 50 and 54 y.o.	Population
M_55	Male population between 55 and 59 y.o.	Population
M_60	Male population between 60 and 64 y.o.	Population
M_65	Male population between 65 and 69 y.o.	Population
M_70	Male population between 70 and 74 y.o.	Population
M_75	Male population between 75 and 79 y.o.	Population
M_80	Male population older than 80 y.o.	Population
F_0	Female population between 0 and 12 months	Population
F_1	Female population between 1 and 4 y.o.	Population
F_5	Female population between 5 and 9 y.o.	Population
F_10	Female population between 10 and 14 y.o.	Population
F_15	Female population between 15 and 19 y.o.	Population
F_20	Female population between 20 and 24 y.o.	Population
F_25	Female population between 25 and 29 y.o.	Population
F_30	Female population between 30 and 34 y.o.	Population
F_35	Female population between 35 and 39 y.o.	Population

F_40	Female population between 40 and 44 y.o.	Population
F_45	Female population between 45 and 49 y.o.	Population
F_50	Female population between 50 and 54 y.o.	Population
F_55	Female population between 55 and 59 y.o.	Population
F_60	Female population between 60 and 64 y.o.	Population
F_65	Female population between 65 and 69 y.o.	Population
F_70	Female population between 70 and 74 y.o.	Population
F_75	Female population between 75 and 79 y.o.	Population
F_80	Female population older than 80 y.o.	Population
child_pop	Child population	Population

Table 3.3: Input data variables with explanation and context.

3.3 Processing

After collecting all the data and aggregating them at the hexagonal level, we pre-process the data through scaling and imputing missing values.

3.3.1 Data scaling

Feature scaling is an important step in data processing, and it is needed to standardize the range of the features of the data. A technique that is robust to outliers is the robust scaler, given by:

$$\frac{x - Q_2(x)}{Q_3(x) - Q_1(x)} \tag{3.1}$$

where Q_i is the i-th quantile. This scaler subtracts the sample median and divides it for the sample interquartile range (IQR). The scaling happens independently on each feature with the statistics computed with the samples of the training set. Median and IQR are then used to scale the test set.

Typically standardization uses the mean and the standard deviation, instead of the median and IQR. However, these two quantities can be negatively influenced by outliers.

3.3.2 Imputing

In Figure 3.5, we show the amount of missing values per dimension and we highlight the importance of not imputing those missing values to not overestimate poverty. Now, we want to investigate how to deal with missing values regarding the input features. An interesting way to impute these values is through k- Nearest Neighbors [66]. For each observation that has missing values, those are obtained computing the mean of k nearest neighbors (that have a value for the missing feature) of the training set. The distance between two observations is computed based on the not-missing features. The default distance used is the Euclidean distance, defined in Equation 3.2.

$$d(x,y) = \left(\sum_{i=1}^{n}(x_i - y_i)^2\right)^{1/2} \quad \text{for } x, y \in \mathbf{R}^n \tag{3.2}$$

The features of the k-Nearest Neighbors (KNN) selected can be uniformly averaged or weighted based on the distance. Let us observe that since for each observation and for each feature we are taking the nearest samples that do not have a missing value for that feature, for the same observation we may have different neighbors for different features, depending on the amount of missing values. If the number of samples with a feature available is less than k and defining distance is not possible, the mean of the training set for that feature is imputed in the remaining samples. While if a feature is missing in the whole training set, it is removed.

However, KNN's cost grows with the size of the data, making it computationally expensive for large datasets. So in that case, a substitute imputer is using the median to fill the missing data.

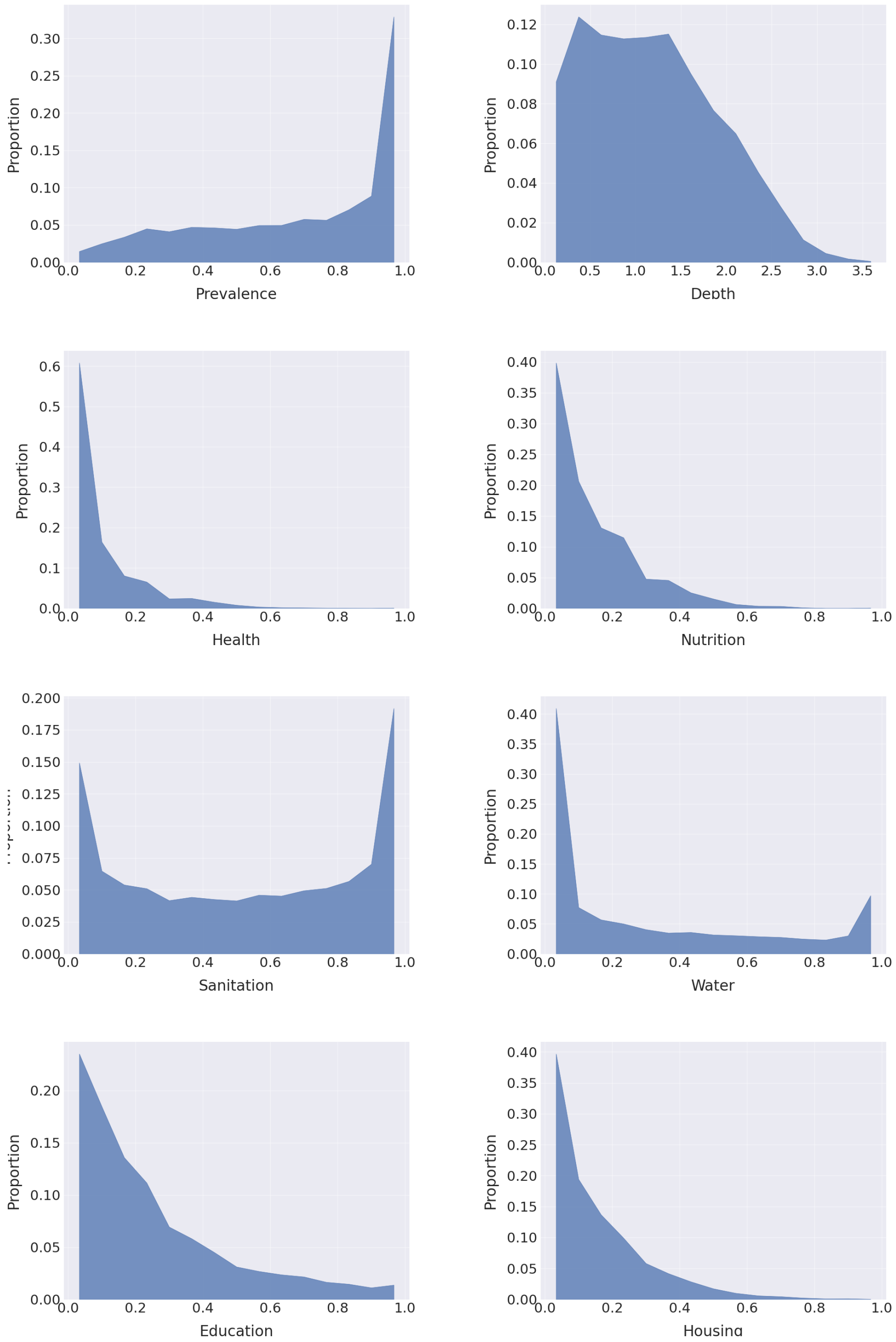


Figure 3.7: Distribution of the output variable for prevalence, 2 or more deprivations, 3 or more deprivations, depth, housing, water, sanitation, nutrition, health and education.

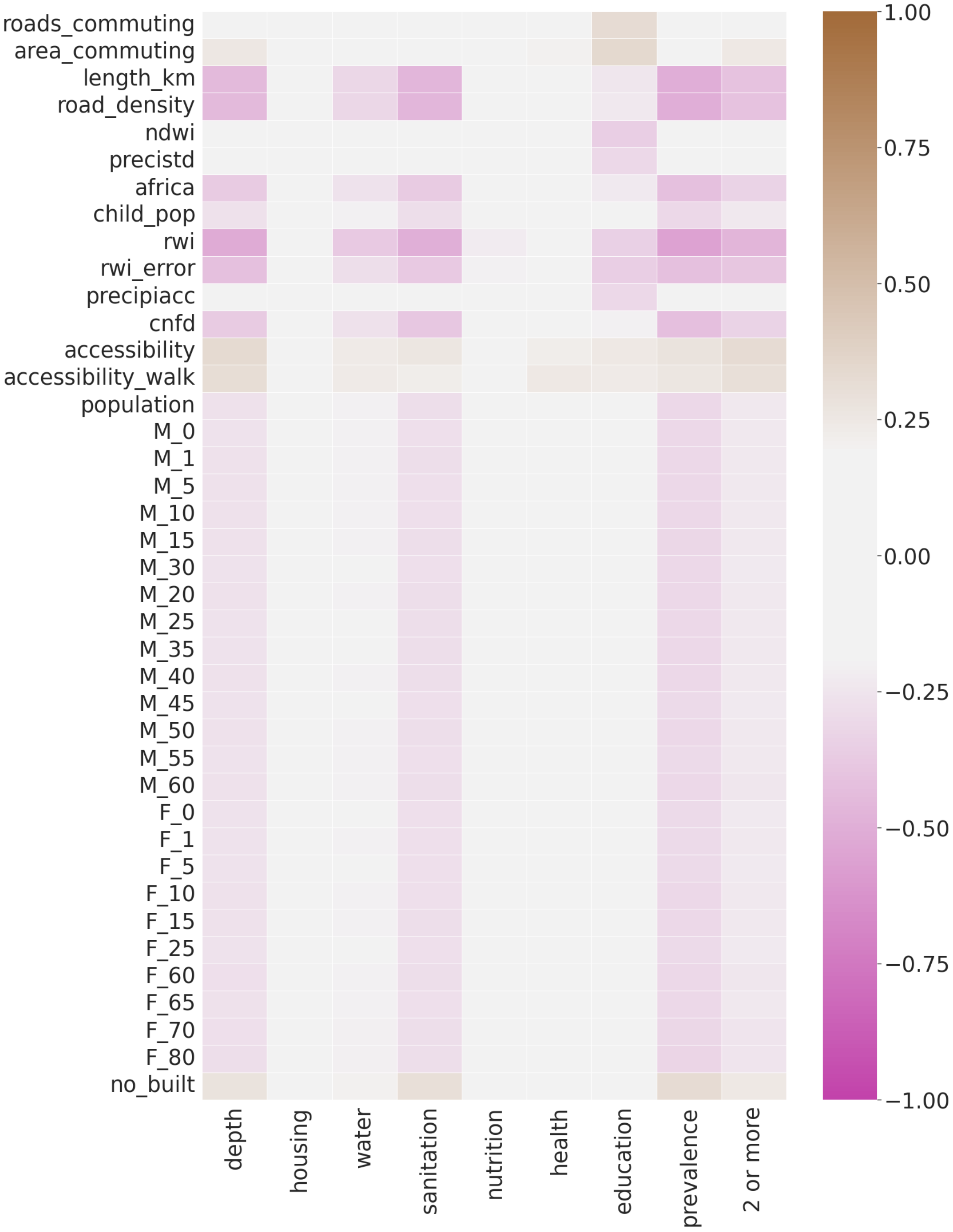


Figure 3.8: Correlation between features and response variables with correlation above absolute value 0.3.

4
Models

Supervised learning is a part of statistical learning where the objective is to establish a relation between a response variable and a set of predictor variables, while unsupervised learning consists in finding patterns in data with no predefined response Y. We can observe the difference between explanatory modeling and predictive modeling. In the first, the focus is to understand the causal relation between X and Y, while in the second, the goal is to predict the response starting from X [67]. In this thesis, we will deal with a supervised learning problem, focusing on predictive modeling.

Hence, we want to learn a function $f : X \rightarrow Y$, called hypothesis function, that outputs $y \in Y$ from $x \in X$. In a supervised learning setting, we have $\mathcal{D} = \{(x_1, y_1), ..., (y_n, x_n)\}$ that is our training set composed by n instances i.i.d. drawn from a joint probability distribution $\mathbf{P}_{X,Y}$. In a statistical learning framework, we want to build a model that can provide accurate predictions on new, unobserved data from $\mathbf{P}_{Y,X}$. Hence, predictive modeling can be seen as a function estimation problem [68], where the accuracy of the function is assessed through a loss function $L(\hat{y}, y)$, that assesses the discrepancy between the the predicted value $\hat{y}$ from the outcome y. Associated to an hypothesis function, we have a risk, that is defined as the expectation of the loss function:

$$R(h) = \mathbf{E}[L(h(x), y)] = \int L(h(x), y) d\mathbf{P}(x, y) \tag{4.1}$$

Thus, the objective of a learning algorithm is finding the hypothesis function $h^* \in \mathcal{H}$ that minimized the risk:

$$h^* = \arg\min_{h \in \mathcal{H}} R(h) \tag{4.2}$$

However, to compute the risk, the joint probability $\mathbf{P}_{X,Y}$ is required, therefore what can be done in practice is compute an approximation of it, called empirical risk, that is an empirical estimate of the true risk of the model, where the expectation of the loss is taken with respect to the empirical distribution $\hat{\mathbf{P}}_{X,Y}$, that assigns $1/n$ to each data point (uniform distribution):

$$R_{\text{emp}}(h) = \frac{1}{n} \sum_{i=1}^{n} L\left(h\left(x_i\right), y_i\right) \tag{4.3}$$

The empirical risk minimization principle [69] states that the learning algorithm should choose $\hat{h} \in \mathcal{H}$ that minimizes the empirical risk:

$$\hat{h} = \arg\min_{h \in \mathcal{H}} R_{\text{emp}}(h) \tag{4.4}$$

By the strong law of large numbers,

$$\lim_{n \to \infty} R_{\text{emp}}(h) = R(h^*), \tag{4.5}$$

Hence, the objective of the learning algorithm is to solve an optimization problem. We need to specify the hypothesis space to make the problem not ill-posed. Restricting the function space defines a class of model, called model class. These restrictions can be seen as a complexity restriction of some kind, since most model classes impose, explicitly or implicitly, some simple structure in small neighborhoods of the input space X.

While we choose a model minimizing the empirical risk, the overall goal is to have the lowest true risk, being able to perform well on new, independent data from $\mathbf{P}_{X,Y}$. Estimating a model with as few assumptions as possible is preferable, but a too flexible model may not generalize well to unseen data. This phenomenon is called overfitting. On the other hand, if the model is not flexible enough to capture the complex structure of the data, it leads to underfitting.

An hypothesis function with low expected risk $\mathbf{E}[R(\hat{h})]$ tends to generalize well. Specific to the squared error loss, this quantity can be decomposed in terms of bias and variance. We can visualize the trade-off between variance and bias in Figure 4.1.

- Bias: bias errors are caused by erroneous assumptions about the model space. High bias can cause underfitting, since the model does not capture relevant relations between the

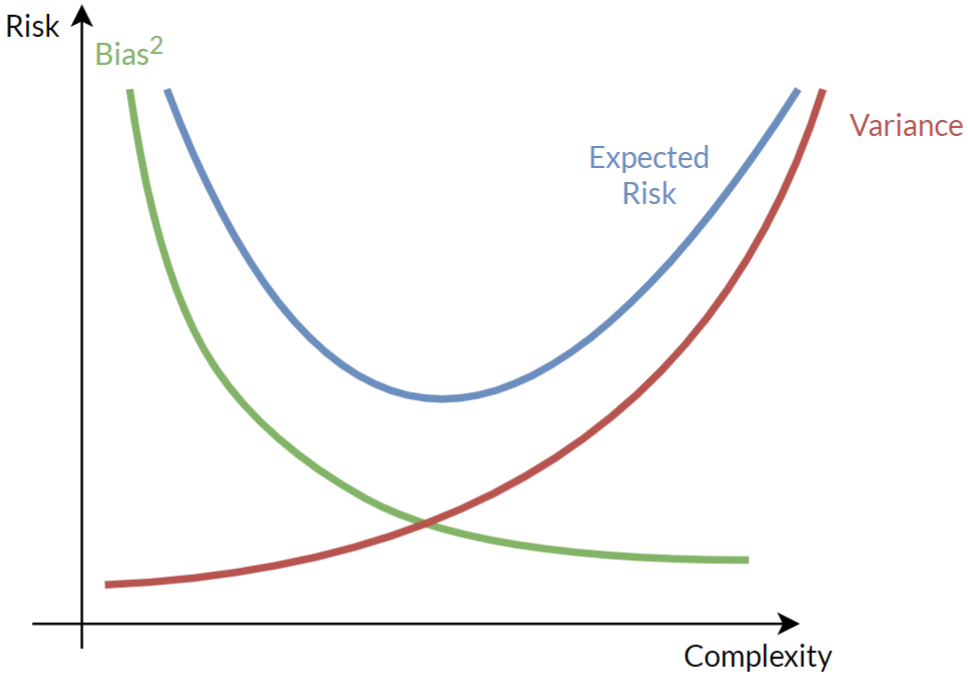


Figure 4.1: Graph of the trade-off between variance and bias. A model with high complexity has high variance overfits the data and it is not able to generalize well. While a model not complex enough cannot capture the patterns in the data, and it has high errors in the training and test data.

predictors and the response variable. The bias of an estimator $\hat{h} \in \mathcal{H}$ is the difference of the expected value of the estimator at that point and the the true value:

$$\text{Bias}(\hat{h}(z)) = \text{E}[\hat{h}(z)] - f(z) \tag{4.6}$$

- Variance: the model is too sensitive to small fluctuations in the training set. High variance may lead to overfitting and the functions also model random noise.

$$\text{Var}(Z) = \text{E}\left[(Z - \text{E})^2\right] = \text{E}[Z^2] - \text{E}[Z]^2 \tag{4.7}$$

We can write the expectation of the conditional risk at x for the squared loss:

$$\begin{aligned} \text{E}[R(\hat{h}(x))] &= \text{E}\left[(Y - \hat{h}(x))^2 \mid X = x\right] \\ &= \textit{Noise} + \text{Bias}[\hat{h}(x)]^2 + \text{Var}[\hat{h}(x)]. \end{aligned} \tag{4.8}$$

The first term is a noise term, while the other two remaining terms are bias and variance. This decomposition can be done for x. Hence, from Equation 4.8, we can derive the concept of bias-variance trade-off, that highlights the importance of selecting a model balancing its complexity. While the mathematical decomposition holds for the squared error loss, the complexity trade-off can be generalized also for the other losses [70].

4.1 Decision Trees

A decision tree is a non-parametric supervised learning method that can be used for classification or regression [71]. The output is inferred by learning simple decision rules inferred from the features. It can be seen as a piecewise constant approximation [72].

- The advantages of the model is that it is simple to understand and to interpret, and the reasoning for the predictions can be explicitly visualized, as can be seen in Figure 4.2. They can deal with continuous and categorical data and are invariant under monotone transformation of the input. Decision trees capture non-linear relations between features and response, and they perform variable selection. Moreover, the cost of predictions is logarithmic in the number of training data.

- The disadvantages of this model is that over-complex trees do not generalize well, leading to overfitting (that can be mitigated with pruning). Small variation in the data can lead to major changes in the structure of the tree and they have high variance that can be managed through bagging or boosting. Their predictive performance is usually limited and they lack smoothness.

Learning an optimal decision tree is an NP-complete problem [74]. Hence, in practice heuristic algorithms such as a greedy approach, make local optimal decisions at each node. However, these algorithms do not guarantee a globally optimal decision tree, and this can be mitigated by training multiple trees in an ensemble learner. A decision tree can be constructed in the following way: at the root node the best feature is selected and the training data are divided based on a threshold. For the other nodes the feature and the threshold are selected based on the training data that would reach that node. This procedure is repeated in a recursive manner. Given the training data $\{x_i\}_{i=1}^{m} \subset \mathbf{R}^n$, and a output vector $y \in \mathbf{R}^m$, a decision tree recursively partition the feature space so that samples with similar target values are grouped together. Let the data at the node k be represented by D_k with $|D_k| = n_k$. Each split $\theta = (\varphi, t_k)$ is defined by a feature φ and by a threshold t_k, and θ partition D_k in $D_k^{left}(\theta)$ and $D_k^{right}(\theta)$:

$$\begin{aligned} D_k^{left}(\theta) &= \{(x, y) \in \mathbf{R}^n \times \mathbf{R}^m | x_j \leq t_k\} \\ D_k^{right}(\theta) &= D_k \backslash D_k^{left}(\theta) \end{aligned} \tag{4.9}$$

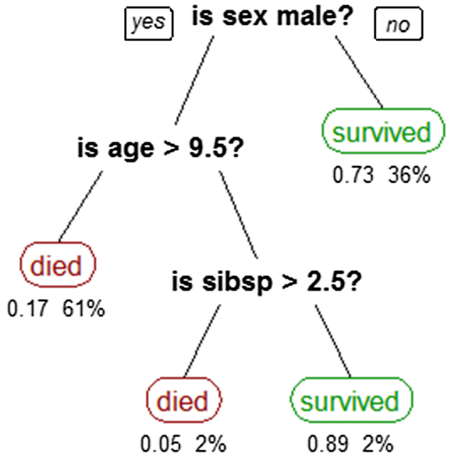


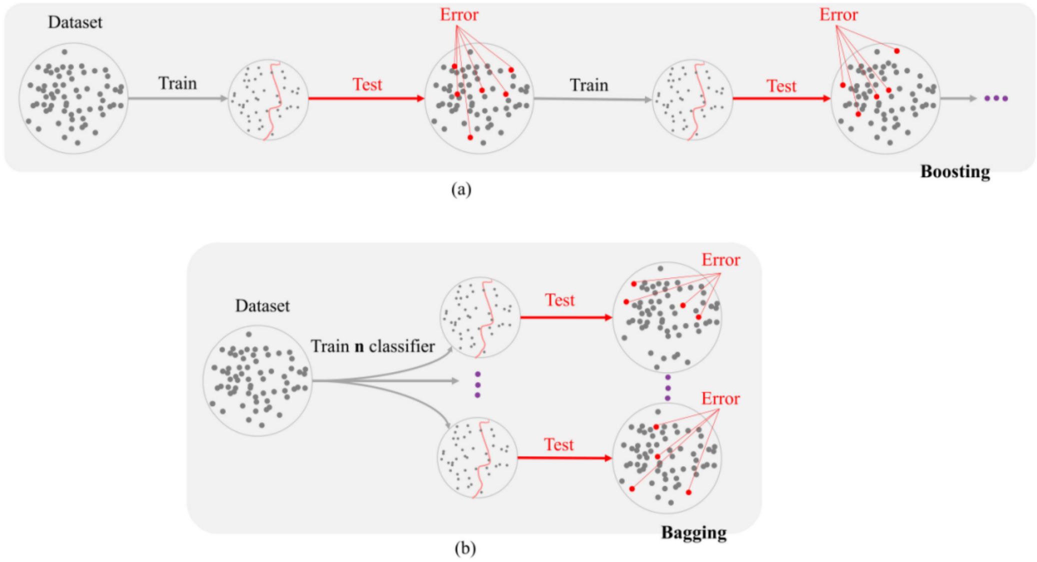
Figure 4.2: Example of a decision tree of the titanic dataset. The data are first split based on the sex feature, and then on age and lastly on the number of siblings and spouses aboard. Here pruning has been applied. We can observe that the decisions taken by the model can be intuitively explainable, making decision trees a very interpretable class of models. [73].

4.2 Ensemble Methods

Ensemble methods are methods that combine a group of weak learners to collectively make a final prediction. A weak learner is a learner whose performance is slightly better than random chance. A single model may not perform so well by itself due to high bias or variance, however, when aggregated it yields a better performance. There are two main types of ensemble learning methods: boosting [75, 76, 77, 78] and bagging [79], and their differences can be visualized in 4.3. Both methods through resampling have different training sets for each classifier. Let us note that combining multiple learners is useful if there is disagreement among them, since if they have identical outputs there is no gain. An ideal ensemble has highly correct classifiers that disagree as much as possible [80, 81].

- In bagging the weak learners are trained in parallel. Each classifier has a different training set that is generated by randomly sampling with replacement the training set. Bagging is effective with “unstable” methods, where a small change in the data lead to larger changes in the predictions, like decision trees. A method that uses bagging is random forest.

- In Boosting, the weak learners are trained sequentially. The training set of each class is chosen based on the performance of the previous learners. At each step the training set is built by sampling with higher probability the observation wrongly predicted at the previous step.



(a)

(b)

Figure 4.3: (a) Boosting: the models are sequentially built on the errors of the previous model, focusing on those predictions wrongly predicted. (b) Bagging: the models are fitted and evaluated in parallel. Image from [82].

4.3 Gradient Boosting

The gradient boosting framework was developed by Friedman and called GBM. It casts the problem as a numerical optimization problem where the objective is to minimize the loss of the model by adding weak learners using a gradient descent like procedure. We can visualize the additive procedure in Figure 4.4. This class of algorithms is described as a stagewise additive model: when a weak learner is added, the existing ones are frozen and left unchanged. This differs from a stepwise approach, that readjusts previously entered terms when new ones are added. Examples of the former approach are XGBoost and LightGBM, while of the latter is Adaptive boosting (AdaBoost), that iteratively identify wrongly predicted points and it adjust their weights, continuing to optimize in a sequential matter until it obtains the strongest predictor [75].

There are three main components in a gradient boosting algorithm: a loss function to optimize, a weak learner and an additive model [83]. The first two depend on the type of problem to solve and there are different options that will be later listed. The last component is an additive model, where trees are added one at the time keeping the existing one fixed. Gradient boosting algorithm is a iterative functional gradient descent algorithms [84, 85]. At each step,

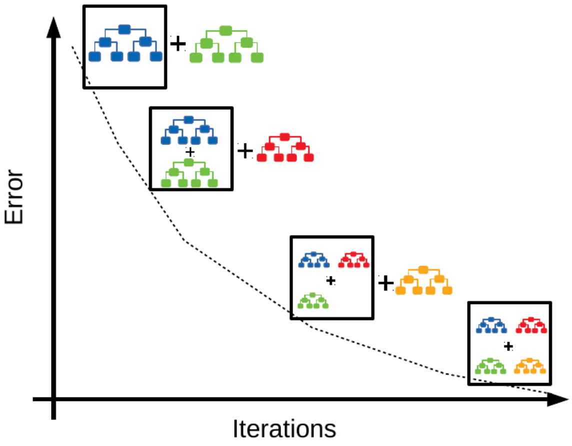


Figure 4.4: As the number of fitted trees additively increases to deal with the wrong predictions of the previous model, the error decreases.

a new base-learner $\phi(x)$ is added to the ensemble:

$$f^{n+1}(x) = f^{n}(x) + \phi(x) = y \tag{4.10}$$

The function ϕ tries to learn the residual of the previous model f^{n}

$$\phi(x) = y - f^{n}(x) \tag{4.11}$$

Through gradient descent the model is trained to minimize any loss function ($y - f(x)$ is the derivative of the mean square error). The new model gives the steepest descent in the loss function, and this is the meaning of the 'gradient' part in gradient boosting. The complete algorithm is shown in Algorithm 4.1.

As previously said, the type of loss and type of learner can vary depending on the problem, that can be a regression, classification or other. We will focus for regression loss, since the problem we want to solve is a regression problem. We can visualize the regression losses in Figure 4.5.

1. Regression :continuous variable $y \in \mathbf{R}$

    - Gaussian $L2$ loss:

$$L_2(y, f(x)) = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

Algorithm 4.1 Gradient Boosting

```
Input Data set $\mathcal{D} = \{(x_i, y_i)_{i=1}^{p}\} \subset \mathbf{R}^n \times \mathbf{R}$
      A loss function $L(\cdot, \cdot) : \mathbf{R} \to \mathbf{R}^+$
      A base learner $\phi \in \Phi$
      The number of iterations $M$.
      The learning rate $\eta$.

Initialize $\hat{f}^{(0)}$ with a constant: $\hat{f}^{(0)}(x) = \hat{f}_0(x) = \hat{\theta}_0 = \arg\min_{\theta} \sum_{i=1}^{n} L(y_i, \theta)$

for $m = 1$ to M
    Compute the negative gradient $\hat{g}_m(x_i) = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=\hat{f}^{(m-1)}(x)}$
    Fit a new base-learner function $\hat{\phi}_m = \arg\min_{\phi \in \Phi, \beta} \sum_{i=1}^{n} \left[\left(-\hat{g}_m(x_i)\right) - \beta\phi(x_i)\right]^2$
    Compute step-size $\hat{\rho}_m = \arg\min_{\rho} \sum_{i=1}^{n} L\left(y_i, \hat{f}^{(m-1)}(x_i) + \rho\hat{\phi}_m(x_i)\right)$
    Shrink learning rate $\hat{f}_m(x) = \eta\hat{\rho}_m\hat{\phi}_m(x)$
    Update function estimate $\hat{f}^{(m)}(x) = \hat{f}^{(m-1)}(x) + \hat{f}_m(x)$
end for

Output $\hat{f}(x) \equiv \hat{f}^{(M)}(x) = \sum_{m=0}^{M} \hat{f}_m(x)$
```

- Laplace $L1$ loss function:

$$L_1(y, f(x)) = \sum_{i=1}^{n} |y_i - f(x_i)|$$

- Huber loss function, δ specified:

$$L_{\text{huber}}(y, f(x) \mid \delta) = \begin{cases} \frac{1}{2}(y - f(x))^2 & \text{for } |y - f(x)| \leq \delta \\ \delta \cdot \left(|y - f(x)| - \frac{1}{2}\delta\right), & \text{otherwise} \end{cases}$$

- Quantile loss function, α specified:

$$L_{\text{quantile}}\left((y_i - f(x_i)) \mid \alpha\right) = \begin{cases} \alpha(y_i - f(x_i)) & \text{if } (y_i - f(x_i)) \geq 0 \\ (\alpha - 1)(y_i - f(x_i)) & \text{if } (y_i - f(x_i)) < 0 \end{cases}$$

2. Classification: categorical variable y

   - Binomial loss function
   - Exponential loss (used in AdaBoost)

3. Other

   - Loss functions for survival models
   - Loss functions counts data (Poisson loss)
   - Custom loss functions

There are several options for the weak learners that can be classified in linear models, smooth models and decision trees. Other types of models, such as Markov random fields [86] or wavelets [87] are used only for specific tasks.

- Linear models: ordinary linear regression, Ridge penalized linear regression, random effect
- Smooth models: P-splines, Radial basis functions
- Decision trees: decision tree stumps, decision trees with arbitrary interaction depth

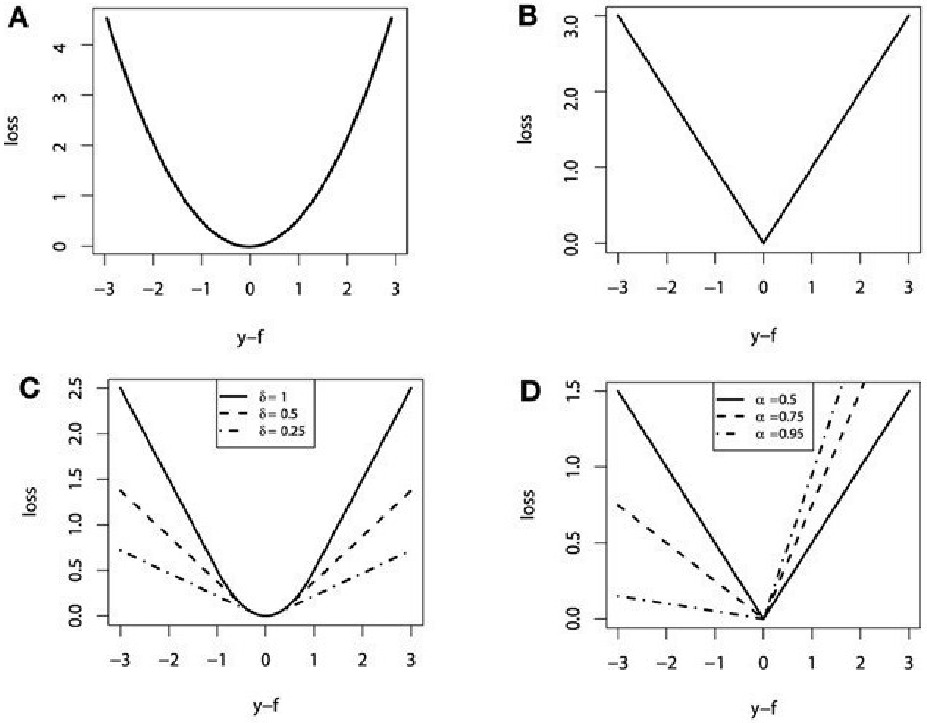


Figure 4.5: Graph representation of the regression losses: (A) $L2$ squared loss function, (B) $L1$ absolute loss function, (C) Huber loss function, (D) Quantile loss function. Images from [83].

- Other models: Markov Random Fields, wavelets, custom base-learner functions

4.4 XGBoost

Extreme Gradient Boosting (XGBoost) is an implementation of gradient boosting decision tree (GBDT) [88], that differs from gradient boosting in the implementation details. XGBoost introduces regularization techniques to control the complexity of the trees to achieve better performance.

$$obj = \sum_{i=1}^{n} L(f(x_i), y_i) + \sum_{k=1}^{K} \Omega(f^{(k)}) \tag{4.12}$$

The first addendum of Equation 4.12 is the loss term, while the second is a regularization term that penalizes complexity of the model [88]. To understand the $\Omega(\cdot)$ function, let us define a tree $f(x)$ such as:

$$f(x) = w_{q(x)}, w \in R^T, \ q : R^p \to \{1, 2, \cdots, T\} \tag{4.13}$$

where w is the vector of scores on leaves, q is a function assigning each point to the corresponding leaf and T is the number of leaves. With these notations, we can define the complexity score of a tree

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|_2^2 \tag{4.14}$$

where γ is a hyperparameter known as complexity parameter. Another technique used to avoid overfitting is using a shrinkage parameter η, that scales the feature weights. Also row subsampling and column subsampling are supported by XGBoost.

Furthermore, other advantages of XGBoost is that it can handle missing values automatically, it allows parallel processing for efficiency, and incremental training, so that the training can be executed in different moments. XGBoost includes a randomization parameter to reduce the correlation between trees, which is important in ensemble models. Lastly, XGBoost employs Newton Boosting, which can be interpret as a Newton method in the function space.

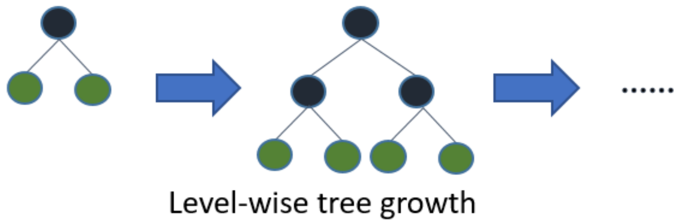


Figure 4.6: XGBoost split all the nodes at a given depth before starting splitting deeper leaves. This approach differs from the LightGBM one. Image from [89].

4.5 LightGBM

LightGBM is a free and open source GBDT framework, developed by Microsoft [90].

The advantages are sparse optimization, parallel training, multiple loss functions, regularization, bagging and early stopping. An important step of LightGBM is the construction of the trees: they do not grow level-wise (row by row), but leaf-wise (best-first) [91]: choosing the leaf that will yield the largest decrease in loss. A visualization of the two techniques can be seen in Figures 4.6 and 4.7

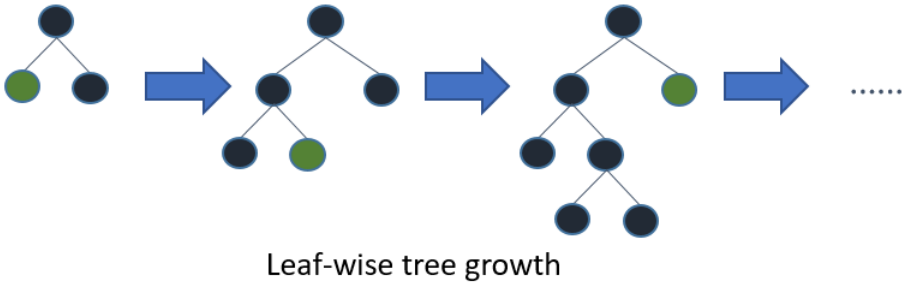


Figure 4.7: LightGBM grows trees in a best-first way, prioritizing the nodes to split, choosing the leaves that maximize the most the loss. Keeping fixed the number of leaves, these leaf-wise growth algorithms tend to achieve lower losses than the level-wise ones. Image from [89].

LightGBM implements a highly optimized histogram-based algorithm that has advantages in terms of efficiency and memory. Two other techniques to increase efficiency are implemented by LightGBM that are Gradient-Based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB).

GOSS is the method to take into account that there are no native weights for data samples in GBDT. Hence, only instances with larger gradients are kept (larger than a predefined threshold or among the top percentiles), since they are the one that contribute the most to information gain.

EFB is a technique to reduce the number of effective features used. In real applications, even if there is a large number of features, the feature space is quite sparse. This fact contributes to designing a nearly lossless method to bundle together nearly exclusive features that rarely take nonzero values simultaneously. Hence, the dimensionality of the feature space is reduced, speeding up the method without hurting accuracy. An example of this is one-hot encoding features. In EFB, we need to specify which features to bundle together and how to construct the bundle. The first issue is NP-hard, so an exact solution cannot be found in polynomial time. Hence, we aim for a good approximation algorithm, reducing the optimal bundling problem to a graph-coloring problem [92], having as vertices the features, and adding edges if the features are not mutually exclusive (or at least allowing a small number of conflicts). The edges are weighted by the number of conflicts between features. Then a greedy algorithm with a constant approximation ratio can produce good results. Then we merge the features in feature bundles, adding offsets to the original features so that they reside in different bins, and the original features can be identified also in the feature bundles.

5
Experiments

In Figure 5.1, we observe the methodology followed to obtain the predictions. The steps in yellow have been explained in Chapter 3.

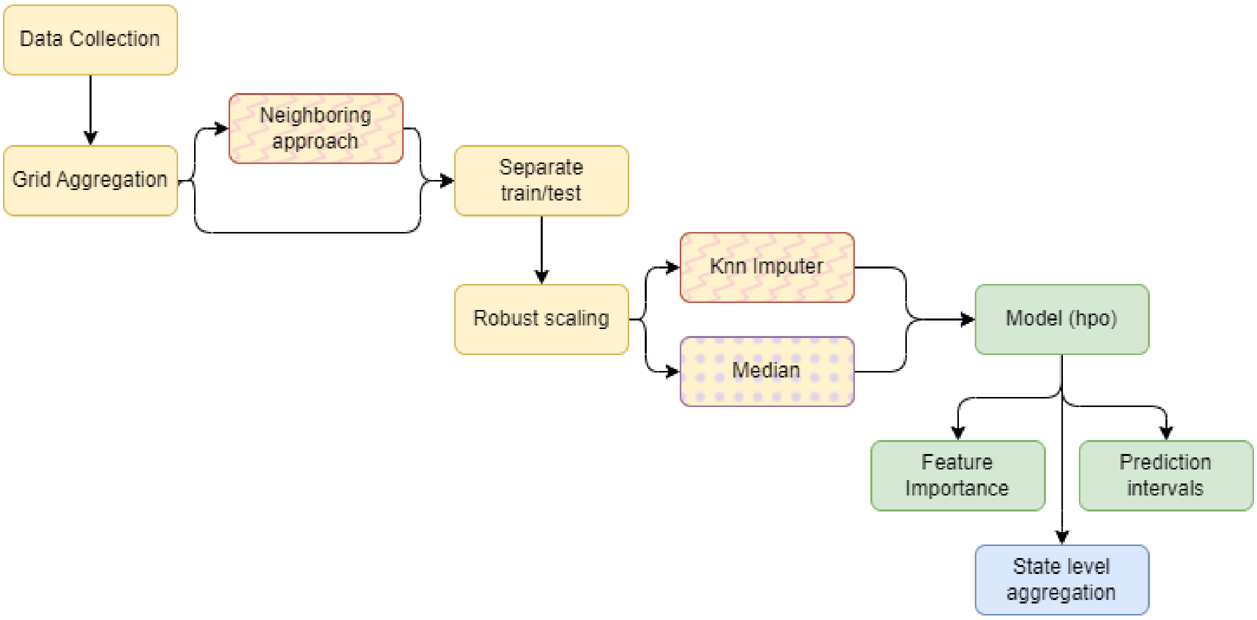


Figure 5.1: Diagram of the methodology pipeline. Yellow represents data collection and processing, green the modeling section and blue validation.The red zigzag represents the steps exclusive for the modeling country per country, while the purple dots represents those for a generalizable model.

5.1 Metrics

The problem is a regression problem, and the outputs are bounded between 0 and 1, with the exception of depth that is between 0 and 4. As metrics, we will consider the mean square error and the R^2.

The mean square error (MSE) measures the average of the squares of the errors. For the empirical risk minimization principle, we can use the MSE as empirical risk, that is the average loss on the observed data set, as an estimation of the true MSE, that is computed on the actual population distribution.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 \tag{5.1}$$

The mean square error can be used to assess the quality of a model function. It is used for regression problems, computing the mean distance from each point to the predicted one. In the definition, the square part is critical to not include negative signs. A lower MSE indicates that the model is closer to the actual data, producing a more accurate model. The disadvantage of the metric is that it heavily weighs outliers.

Another very interesting metric in regression problems is the coefficient of determination, also called R^2, that quantifies the proportion of the variation of the outcome predictable through the input features. Identifying with $\bar{y}$ the mean of the observed data:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{5.2}$$

we define two quantities: the sum of squares of residuals SSR, and the total of squares SST.

$$SSR = \sum_{i=1}^{n}(y_i - f(x_i))^2 \tag{5.3}$$

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{5.4}$$

Hence, we can now define R^2 through these two quantities:

$$R^2 = 1 - \frac{SSR}{SST} \tag{5.5}$$

If the predicted values exactly match with the observed ones, then $SSR = 0$ and $R^2 = 1$.

While a baseline model that always predicts the mean of the observed values results in $SSR = SS_{\text{tot}}$, that gives $R^2 = 0$. However, models that have worse predictions that the baseline have a negative R^2. Therefore, the coefficient of determination can be more intuitively informative than other regression evaluation metrics, whose range vary.

In the Equation 5.5, we can see that the last term is the fraction of the variance unexplained (FVU), since it compares the variance of the model's error with the total variance of the data.

$$R^2 = 1 - \text{FVU} \tag{5.6}$$

5.2 Spatial Cross Validation

If the assumption that the data are i.i.d. does not hold, then randomly splitting the data in train-test does cause data leakage. This fact is often the case with geospatial data. The first law of geography states that "everything is related to everything else, but near things are more related than distant things" [93]. Features are more likely to be similar to the ones in adjacent areas than to distant ones.

Spatial autocorrelation is used to prevent overfitting. However, this may be beneficial in some cases, for example if we want to fill spatial gaps between the training data. While if we want a more generalizable model, spatial autocorrelation would lead to inflating the training data accuracy of a potentially poor model. This could be concerning if this model is then applied to areas where there are no ground truth data to verify the values.

An example of the different spatial folds can be found in Figure 5.2.

5.3 Feature Importance

To have a more transparent and interpretable Machine Learning model we want to interpret the results in terms of overall importance of each feature. SHAP (SHaply Additive exPlanations) is a technique based on cooperative game theory [94], to show the importance of each feature (without evaluating the quality of it).

This approach starts from the shapley values, that are computed averaging the marginal contributions of the predictions across all permutations, in a cooperative game where each feature is a player.

The first is global interpretability, since the collective SHAP values show the positive or negative contribution of each prediction. The second one is local interpretability, where we can

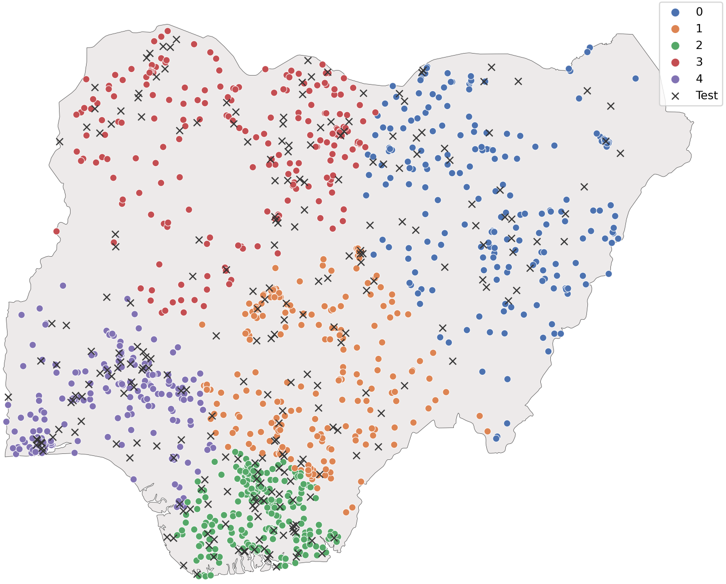


Figure 5.2: Spatial cross validation in Nigeria (for visualization purposes without the neighboring approach. The data are split at random in train and test (where the test size is 20% of the total data), then the training data are splitted in geographically close neighbors in 5 folds.

focus on each observation, understanding the contribution of each feature on an individual case. Lastly, the SHAP values can be computed for any tree-based model. While the SHAP values help provide a more interpretable model, they do not give causality information.

Now we can focus more on the math behind the shapley values. Let us consider a game with N players that collaborate together to obtain a score σ. The contribution of that player is the difference in score if the others would have played without it. To understand how to compute this concept, we can use the following notation: $\Pi = \{\pi_i\}_{i=1}^{N}$ the set of players, $\mathcal{P}(\Pi)$ as the set of parts of Π. Hence, a coalition, that is a combination of player, is an element of $\mathcal{P}(\Pi)$, and we can define the score function as $\sigma : \mathcal{P}(\Pi) \rightarrow \mathbf{R}$. Hence, the shapley value φ_{π_i} of a player π_i can be defined as the weighted sum of the score of games where the player was part of the coalition minus the scores of games where she was not part of the coalition:

$$\varphi_{\pi_i} = \sum_{c \in \mathcal{P}(\Pi) \pi_i \in c} \omega_c(\sigma(c) - \sigma(c \backslash \{\pi_i\})) \tag{5.7}$$

and the weights are given by:

$$\omega_c = -\frac{(|c| - 1)!(N - |c|)!}{N!} \tag{5.8}$$

where $|c|$ is the number of players of the coalition including π_i. This methodology is computationally expensive, because it would mean retraining the model for a large number of possible feature combinations. Hence, we can use an approximation, based on the idea that removing one or more features from the model is approximately equal to compute the expected value of the predictions over all possible values of the removed features. With this approximation, we can more efficiently compute the values, that are the SHAP values.

5.4 Prediction Intervals

In a regression problem, the likelihood of a point prediction of a numeric target is very low, hence including prediction intervals makes the predictions more robust. Predictions uncertainty can be caused by lack of data quality, like noise, or quantity issues, where the true complex distribution is not captured. Estimating uncertainty is essential for sensitive domains and it can be used to build trust in decisions taken by the algorithms.

A prediction interval is an estimate of an interval in which a future observation will fall, with a certain probability, given what has already been observed. Prediction intervals are not the same as confidence intervals. While a confidence interval pertains to a statistic estimated from multiple values and expresses sampling uncertainty, a prediction interval expresses inherent uncertainty in a particular data point on top of the sampling uncertainty, and is thus wider. There is a trade-off between interval width and tolerance, i.e., the percentage of mistakes permissible. For the purpose of this work, we consider the 95% prediction intervals for each data point.

We use a Model Agnostic Prediction Interval Estimator (MAPIE) to compute the prediction intervals [95]. The estimator includes aleatoric and epistemic uncertainties and it is based on the theory of conformal prediction method [96, 97, 98, 99, 100, 101, 102]. MAPIE is based on cross validation, relying on it to obtain conformity scores on the whole training set and perturbed models. These are then combined to estimate prediction intervals on new data with strong theoretical guarantees.

6
Results

In this chapter we are going to present the results of the experiments. We can divide the experiments carried out in two main sections. In Section 6.1, the focus is interpolating the DHS data to obtain predictions for all the hexagons of that country, so each model will be country specific. In Section 6.2, the objective is creating a model that we apply to all the countries in sub-Saharan Africa, extrapolating the DHS data to the countries that do not have DHS data.

6.1 Model dependent on country

The goal of this experiment is producing predictions for the remaining hexagons for those 25 countries that have DHS surveys. Hence, we employ a separate model for each country and for each dimension. In this experiment we implement the neighboring approach, as a way to augment the data considering the smooth variation with each area. The test size is 20% of the data.

We train two models, an XGBoost and a LightGBM model, for each country and for each dimension, using spatial cross validation to select the best hyperparameters. The AutoML library Flaml, developed by Microsoft, has been employed fixing each model for hyperparameter tuning [103], using a time budget of 10 seconds.. In Figure 6.1, we can compare the average performance of XGBoost and LightGBM for each dimension, averaging the results obtained in each country as they are, and weighted on the test size. We can observe that in almost all dimensions, XGBoost performs better than LightGBM, except for depth, nutrition, and prevalence, where the performance is very similar. So we can conclude that XGBoost better captures the distribution of poverty in this situation, and we will show the next results in terms of this model.

While the average of the R^2 of the different countries summarizes in one value the performance of that dimension, it does not capture its distribution. Hence, we plot in Figure 6.2 the boxplot of the performance of each dimension in all the countries. Here we observe that the proportion of children deprived in at least 4 dimensions has on average a performance of $R^2 > 0.55$, but the country's performance are very spread out, reaching even very low results (close to zero). In this sense nutrition and health are even worse, since the models for some countries have a negative R^2 on the test set, that means that for some countries the model is worse than a constant model. This scores could be explained by the lack of data in these two dimensions, that can be observed in Figure 3.5.

Now we want to disaggregate these results to analyze and compare the performance of each individual country. We select prevalence and depth as the main dimensions and we plot their results in a barplot in Figure 6.3. While the performance of the dimensions seen in Figure 6.1 are all above 0.4, the scores of the countries are very spread out, with the worst R^2 values obtained by Malawi and Rwanda.

We can then focus on the performance of a single country, plotting the predicted values of prevalence and depth with the true ones. We show the results for Nigeria in Figure 6.4.

Finally we display the results for all the countries and all the dimensions in Figure 6.1 and 6.2.

The last part of the experiment is to evaluate how much these models can generalize. Hence, after selecting the hyperparameters in the training set for each country, we use them to retrain a model on all the data of the country (including also the test set), and we test the model separately on each country. The performance is evaluated in terms of R^2, and the results can be seen in Figure 6.5, where negative values of R^2 are cut off at 0.

6.2 Generalizable model

In the previous section, for each country a model was trained focused on that country, and as it can be seen in Figure 6.5, these models do not generalize well. Hence, we use a different strategy, creating a model trained on multiple countries. To do this, we randomly select 5 countries (i.e. the 20% of the countries with DHS information) that we keep separate as part of the test set. The countries considered are Angola, Burundi, Guinea, Sierra Leone and Uganda. The

Country	Prevalence	2 or more	3 or more	4 or more	Depth
AGO	0.64	0.83	0.45	0.58	0.72
BEN	0.51	0.64	0.66	0.45	0.74
BFA	0.53	0.63	0.49	0.87	0.78
BDI	0.59	0.41	0.23	0.02	0.42
CMR	0.72	0.76	0.7	0.75	0.66
COM	0.4	0.27	0.5	0.13	0.42
COD	0.41	0.61	0.27	0.42	0.76
GAB	0.72	0.69	0.92	0.74	0.73
GIN	0.53	0.56	0.62	0.58	0.54
KEN	0.37	0.45	0.46	0.63	0.58
LSO	0.54	0.47	0.47	0.6	0.47
LBR	0.26	0.4	0.45	0.35	0.4
MWI	0.23	0.26	0.23	0.08	0.19
MLI	0.76	0.48	0.4	0.84	0.5
MOZ	0.64	0.58	0.56	0.8	0.62
NAM	0.76	0.59	0.9	0.75	0.72
NER	0.46	0.29	0.32	0.79	0.36
NGA	0.51	0.45	0.52	0.53	0.56
RWA	0.38	0.24	0.38	0.17	0.23
SEN	0.65	0.69	0.77	0.69	0.74
SLE	0.41	0.39	0.22	0.1	0.54
TZA	0.43	0.71	0.48	0.77	0.52
TGO	0.76	0.41	0.38	0.59	0.42
UGA	0.4	0.25	0.35	0.48	0.35
ZMB	0.55	0.45	0.39	0.52	0.4

Table 6.1: R^2 scores of XGBoost model for all countries for depth, prevalence and prevalence of 2, 3, 4 dimensions.

location of these countries can be visualized in Figure 6.6. Thus, the observations of the test set represent 19.8% of the sample size. From here, we follow the methodology shown in Figure 5.1, without the neighboring approach and imputing the missing values through the median, since KNN is computationally expensive on large datasets. Another step to the methodology is encoding the information about which country the hexagon belongs to through one-hot encoding.

Furthermore, we explore how different cross validation techniques can help us to obtain a more generalizable model. In Figure 6.7, we can observe how the data are splitted in different folds with spatial cross validation and with the traditional (random) cross validation. With

Country	Housing	Water	Sanitation	Nutrition	Health	Education
AGO	0.7	0.58	0.87	0.32	0.63	0.51
BEN	0.65	0.51	0.55	0.64	0.76	0.72
BFA	0.8	0.57	0.44	0.73	0.77	0.58
BDI	0.35	0.34	0.59	0.32	0.55	0.63
CMR	0.51	0.74	0.74	0.94	0.77	0.85
COM	0.8	0.57	0.52	0.39	0.27	0.73
COD	0.21	0.43	0.82	0.52	0.89	0.31
GAB	0.35	0.78	0.74	0.87	0.37	0.57
GIN	0.32	0.64	0.5	0.48	0.4	0.55
KEN	0.74	0.33	0.3	-0.85	0.75	0.75
LSO	0.23	0.21	0.36	0.81	0.42	0.42
LBR	0.17	0.31	0.31	0.54	0.61	0.22
MWI	0.49	0.29	0.19	0.35	0.28	0.49
MLI	0.65	0.33	0.7	0.68	0.86	0.5
MOZ	0.8	0.45	0.69	-0.02	-0.26	0.43
NAM	0.92	0.52	0.81	0.17	0.37	0.78
NER	0.36	0.3	0.43	0.55	0.39	0.28
NGA	0.24	0.35	0.51	0.56	0.53	0.74
RWA	0.25	0.35	0.5	0.5	0.09	0.53
SEN	0.36	0.73	0.62	0.62	0.74	0.49
SLE	0.22	0.6	0.55	0.2	-0.0	0.56
TZA	0.4	0.44	0.45	0.79	0.56	0.73
TGO	0.41	0.25	0.63	0.69	0.67	0.76
UGA	0.67	0.46	0.42	-0.45	0.58	0.71
ZMB	0.27	0.42	0.49	0.18	-0.41	0.51

Table 6.2: R^2 scores of XGBoost model for all countries for housing, water, sanitation, nutrition, health and education.

spatial cross validation, we can observe that hexagons of the same fold are geographically close to each other, and the folds are built so that there are almost the same number of elements in each one of them. While with the standard cross validation technique, the elements in the same fold are spread across the whole area.

Following the insights obtained in Figure 6.1, we can see that XGBoost performs better, and hence we employ this model, exploiting as hyperparameter tuner the Flaml AutoML framework, giving 120 seconds as time budget for each dimension. The results can be found in Table 6.3. After training and evaluating the model, we predict poverty in the whole sub-Saharan Africa, and the results can be seen in Figure 6.8 for prevalence and in Figure 6.10 for depth.

Dimension	Spatial cv	Random cv
Prevalence	0.43	0.42
2 or more	0.40	0.40
3 or more	0.28	0.27
4 or more	0.14	0.12
Depth	0.45	0.45
Sanitation	0.33	0.34
Water	0.26	0.24
Housing	0.081	0.060
Health	0.079	0.11
Nutrition	0.044	0.044
Education	0.19	0.16

Table 6.3: Comparison of results of applying spatial cross validation and random cross validation for a generalizable model. Each model is evaluated in terms of R^2 on the test set, and the model used is XGBoost.

To facilitate downstream use of the results, we also include the prediction intervals. The predicted intervals are mapped in terms of their lengths, in Figure 6.9 for prevalence and in 6.11 for depth.

Lastly, we focus on better understanding the choice of the model, to have a less "black-box" model and a more interpretable one. We compute the SHAP values plotted in Figure 6.12. We can observe that the most important feature is `avg_rad`, that represents the nighttime light intensity. This result agrees with the studies found in literature that use night luminosity as a proxy for asset-based poverty [21, 22]. In the plot we can observe that low values of night radiance have a positive impact on the outcome, that means towards higher levels of prevalence. After that, we find the critical infrastructure spatial index (CISI), indicated by `africa`, and we can observe that low values of CISI positively impact the model.

An interesting insight is that the age groups with the most impact are older groups, both male and female (`M_75, F_80, F_65, F_75, M_65`). Commuting areas, that show how people move and interact, are useful in predicting poverty, thus we can observe that the amount of kilometers of roads (`win_roads_km_commuting`) and the area (`area_commuting`) are important. Also `road_density` results important. Other features predicted as important are: geographic information about precipitation, such as the Palmer Drought Severity Index (`pdsi`), evapotranspiration (`evapotrans`) and average precipitation (`precimean`), elevation (`elevation`), vegetation (`ndvi`). Furthermore, we can find economical information such as (`rwi`), cell tower information such as mobile tower for 3G network (`UMTS`) and average signal (`avg_signal`), and lastly access to hospital (`accessibility`), also without motorized vehicles (`accessibility_walking_only`).

The last step is to validate the results. To do so, we compare the predictions with the DHS at a subnational. In the surveys, each observation is included with a weight, so that the weighted aggregation is representative. Hence, we aggregate our results at a subnational level, weighting them on the child population. We also compare MICS surveys that do not share the GPS information. Therefore, we compare those results at a national level. In the countries where we have both DHS and MICS surveys, we use the first since they provide more granular information. We have MICS surveys for: Central African Republic, Côte d'Ivoire, Congo, Ghana, Gambia, Guinea-Bissau, Madagascar, Mauritania, Sao Tome and Principe, eSwatini, Chad, South Africa. The results for prevalence can be seen in Figure 6.13, while those for depth in Figure 6.14.

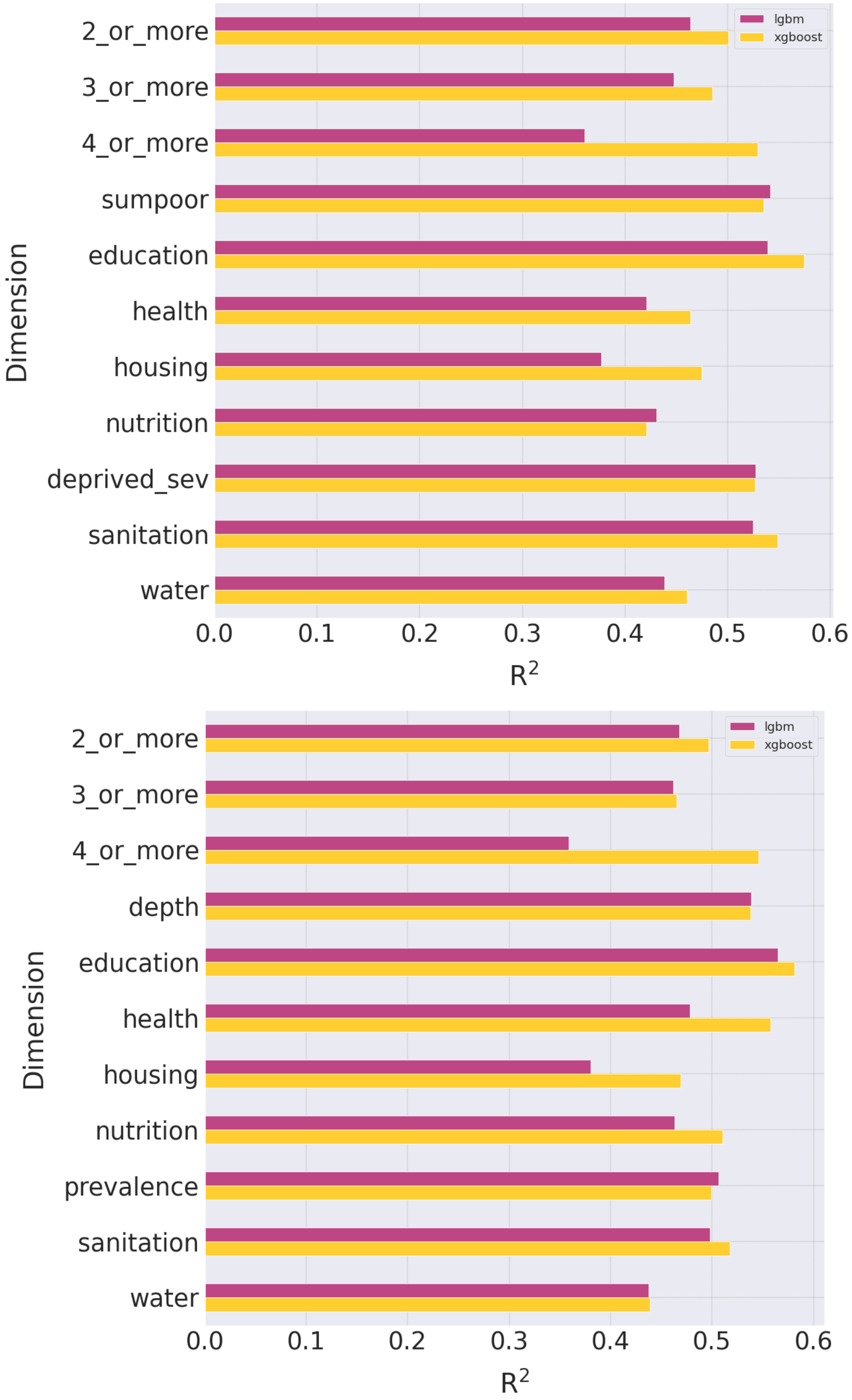


Figure 6.1: Comparison of the performance of XGBoost and LightGMB in terms of R^2 on the test set across different dimensions, averaging the results of the 25 countries with DHS information on the top, and the weighted average of the dimensions on the bottom.

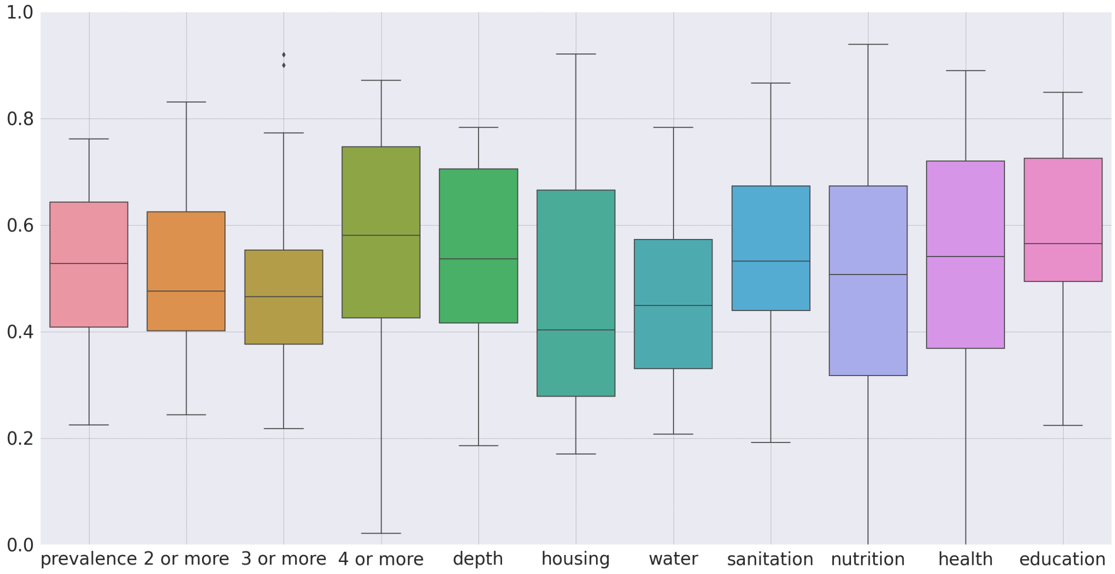


Figure 6.2: Distribution of R^2 values for the different dimensions. We can see that nutrition and health reach really low performance, that can be expected since they have a lot of missing values.

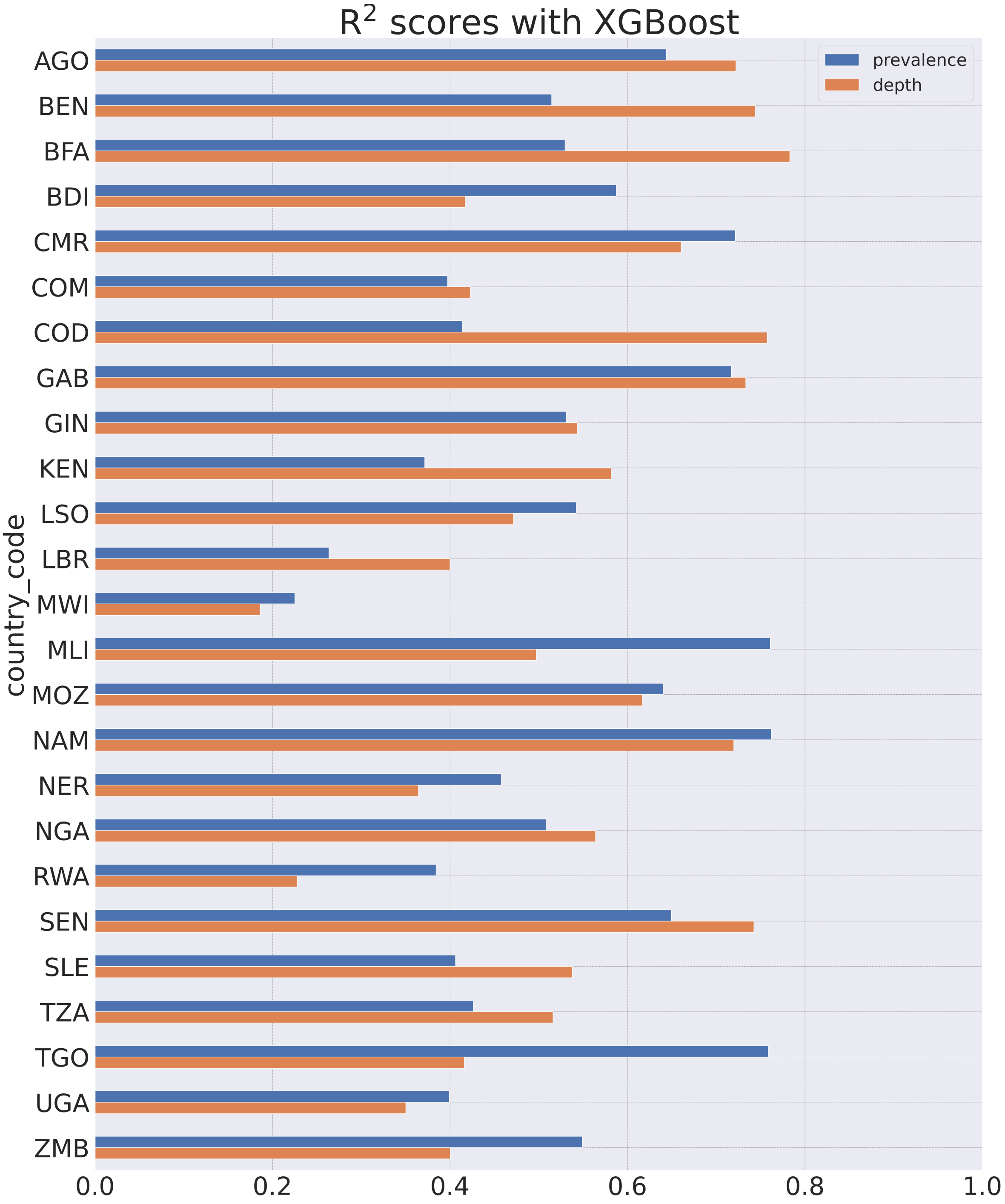


Figure 6.3: Performance of prevalence and depth in terms of R^2 using an XGBoost model per each country. Malawi and Rwanda have the lowest performance, while Namibia is among the best ones.

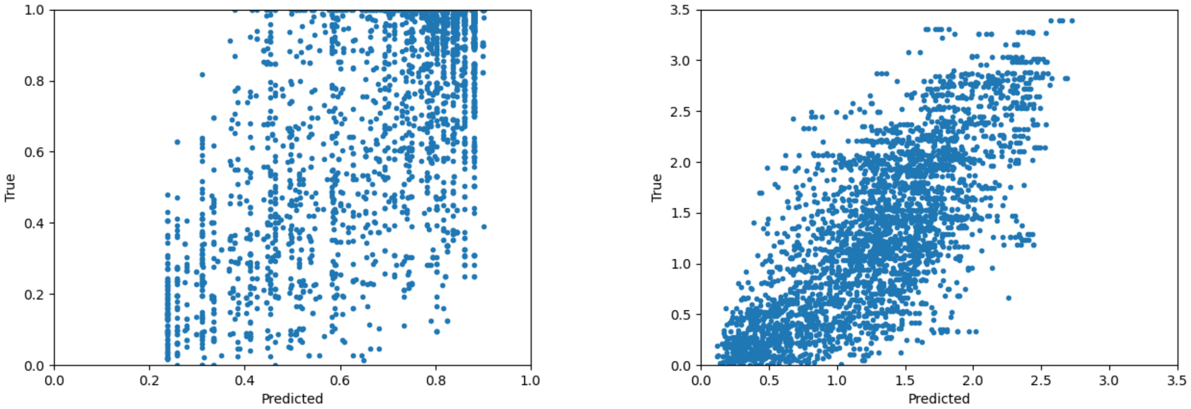


Figure 6.4: The plot compares the predicted outcomes of XGBoost with the true ones. The optimal results would be on the diagonal. The results are for Nigeria and the dimensions plotted are prevalence (on the left) and depth (on the right).

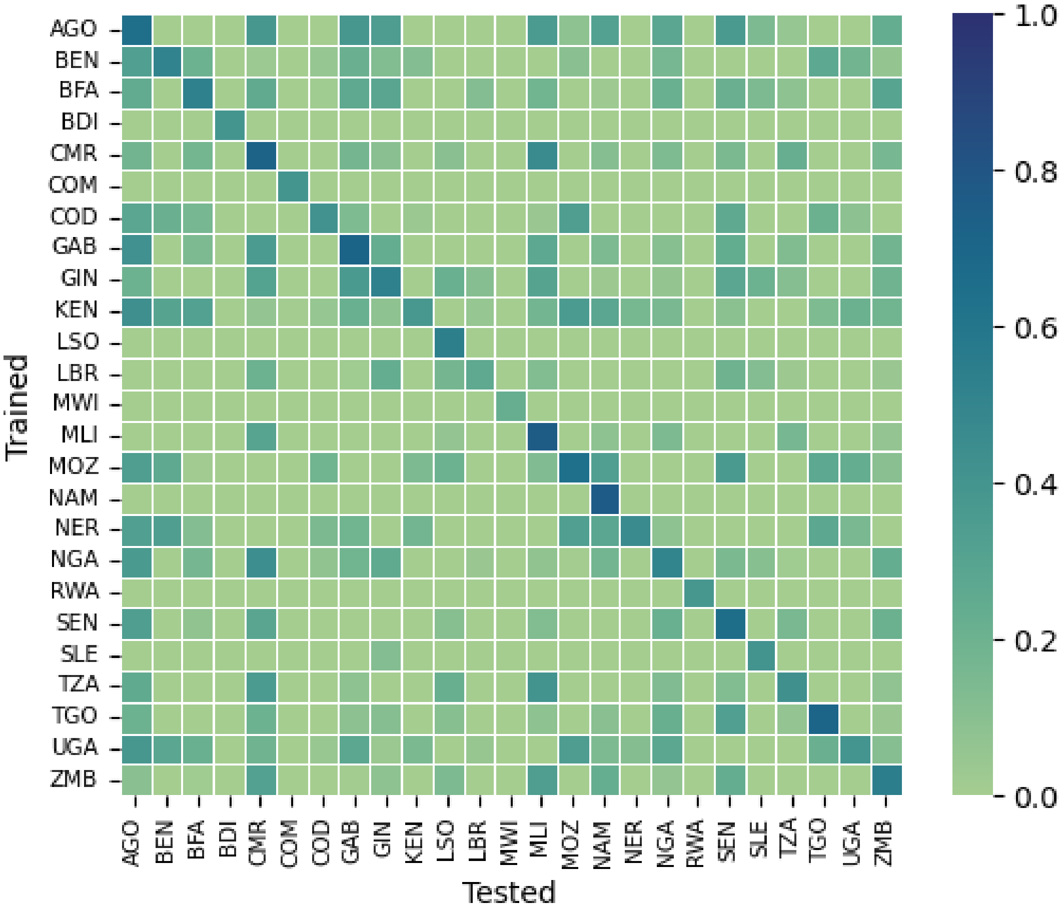


Figure 6.5: For each country, we tested the model trained on one country on the other ones, to see if a similar model could be generalized. The performance are shown in terms of R^2, cutting off at 0. In the diagonal we see the performance of the model on the test of that country.

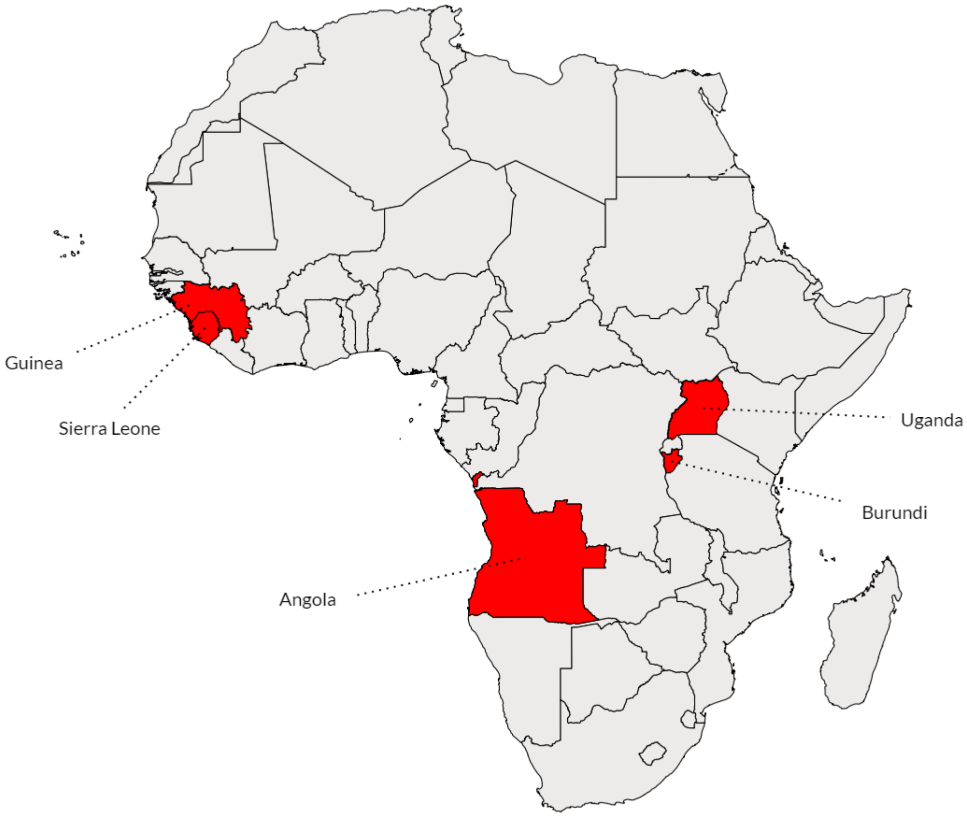


Figure 6.6: Visualization of the countries held out during training as part of the test set: Angola, Burundi, Guinea, Sierra Leone and Uganda.

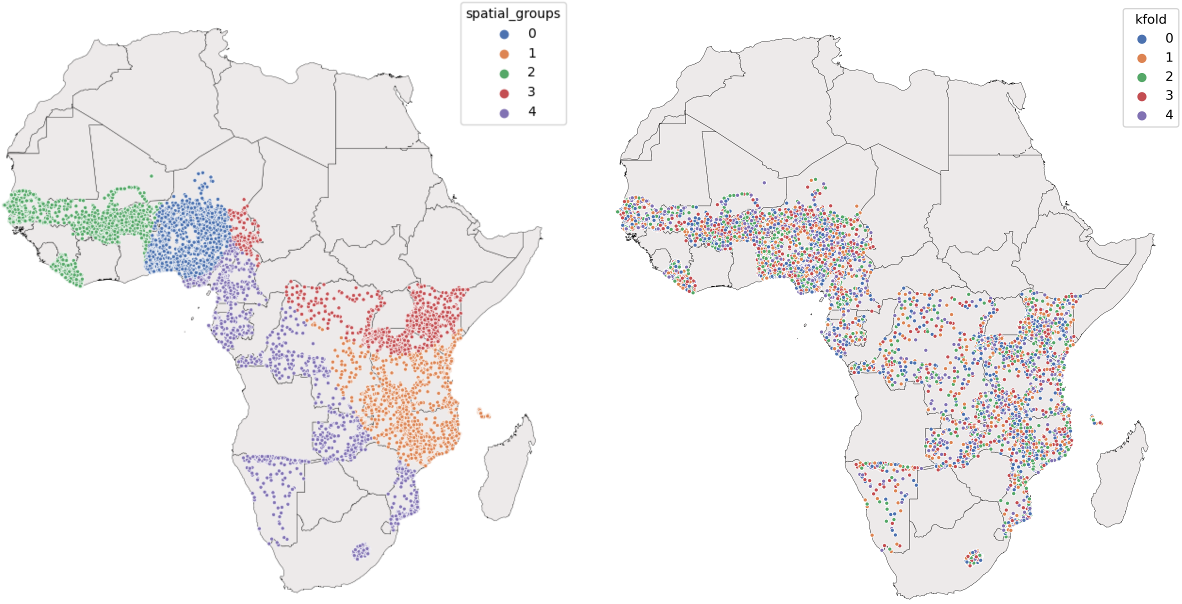


Figure 6.7: On the right a visualization of spatial cross validation is shown, while on the left we can observe the effects of standard cross validation, where the elements are picked at random. The number of folds considered is 5.

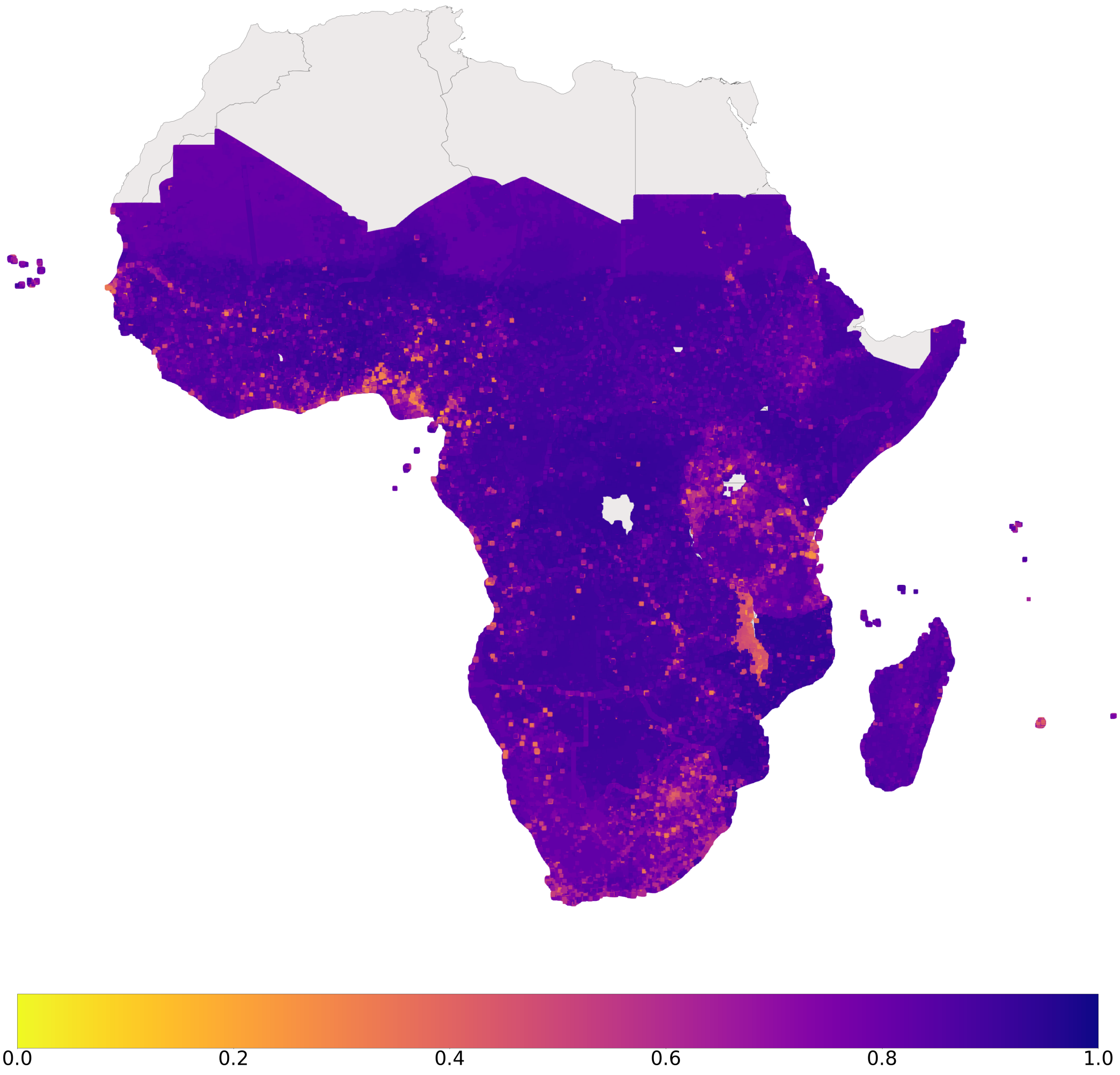


Figure 6.8: Distribution of prevalence in sub-Saharan Africa.

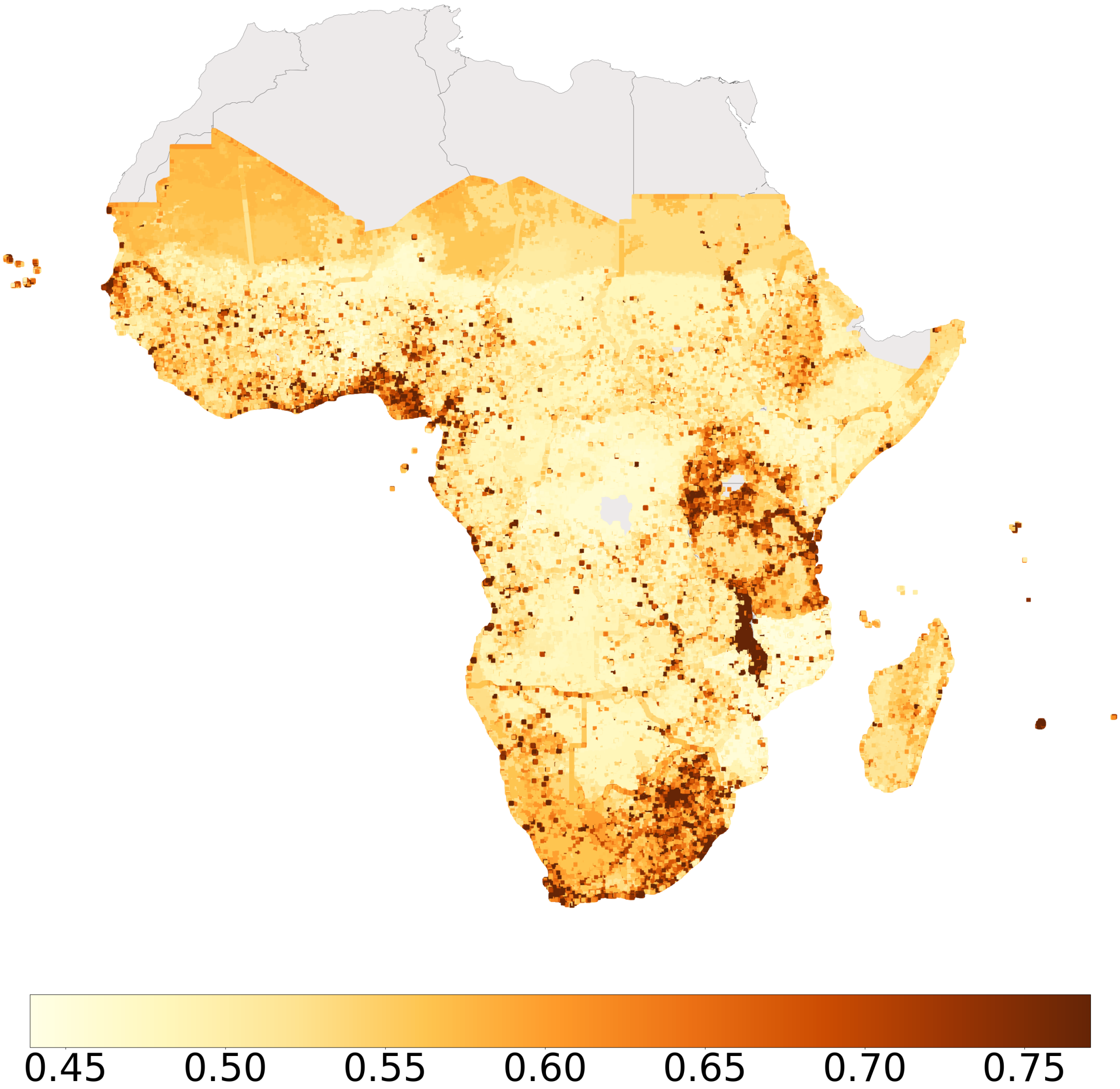


Figure 6.9: Distribution of interval length for prediction intervals for prevalence in sub-Saharan Africa.

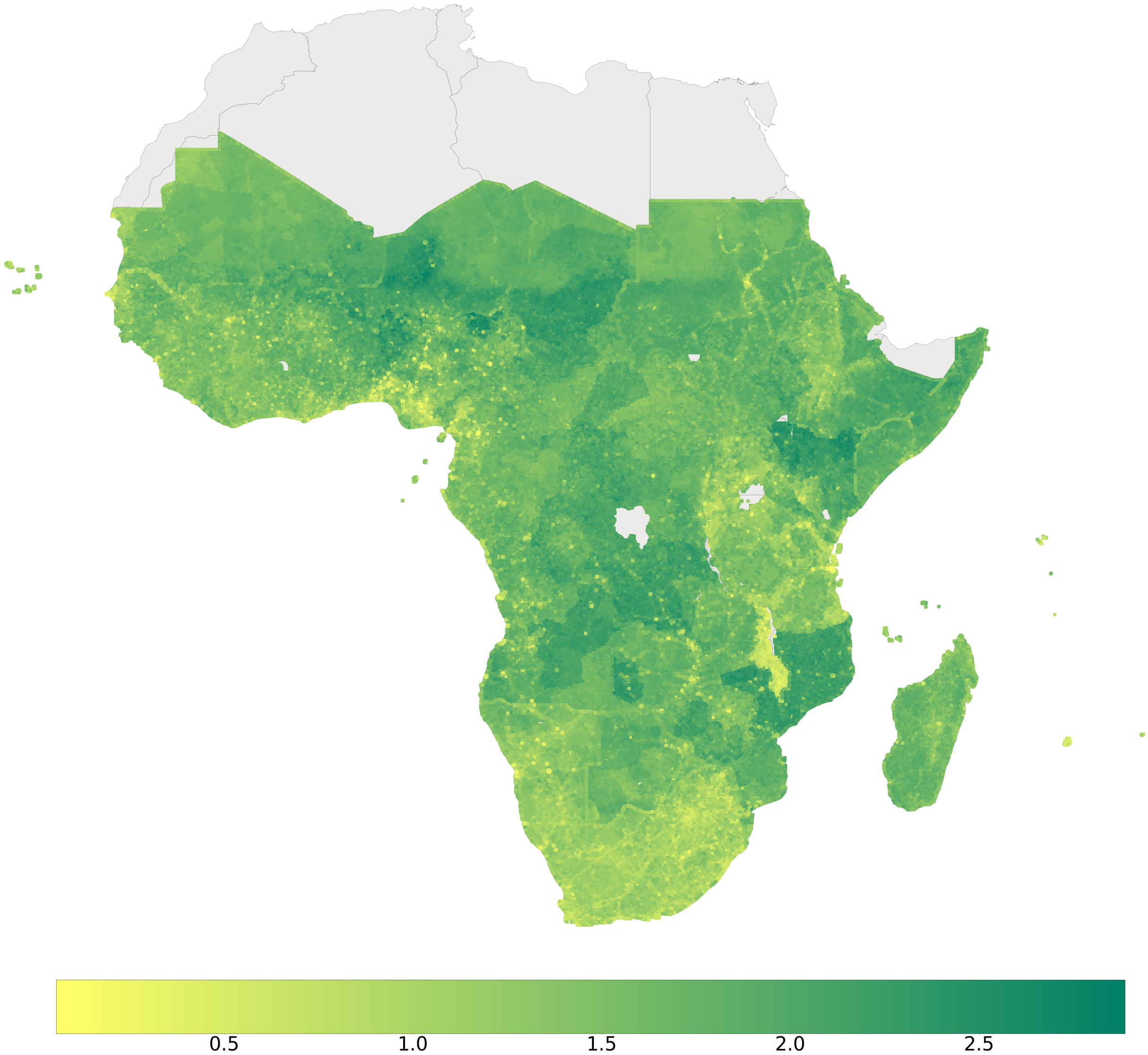


Figure 6.10: Distribution of depth in sub-Saharan Africa.

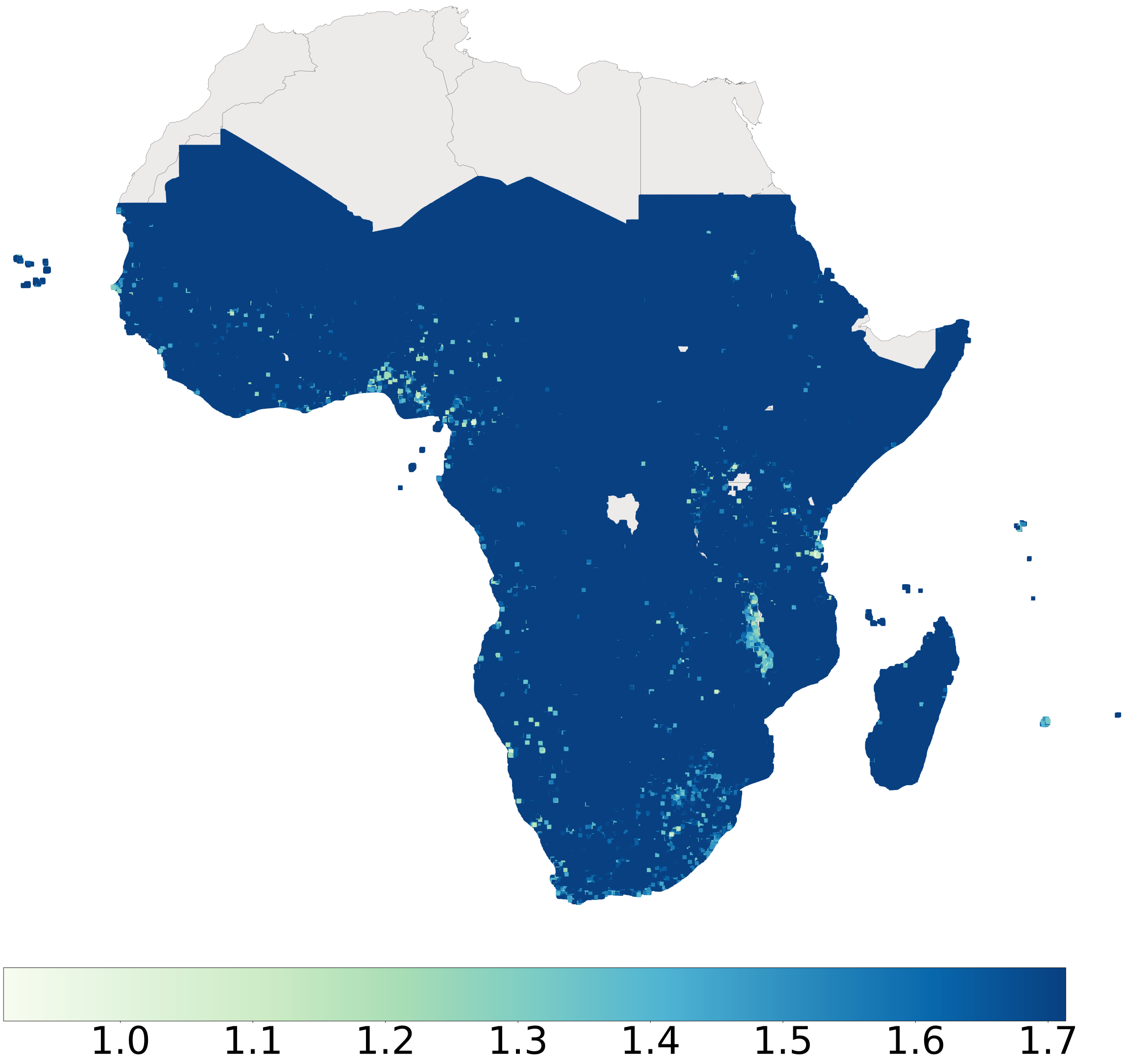


Figure 6.11: Distribution of interval length for prediction intervals for depth in sub-Saharan Africa.

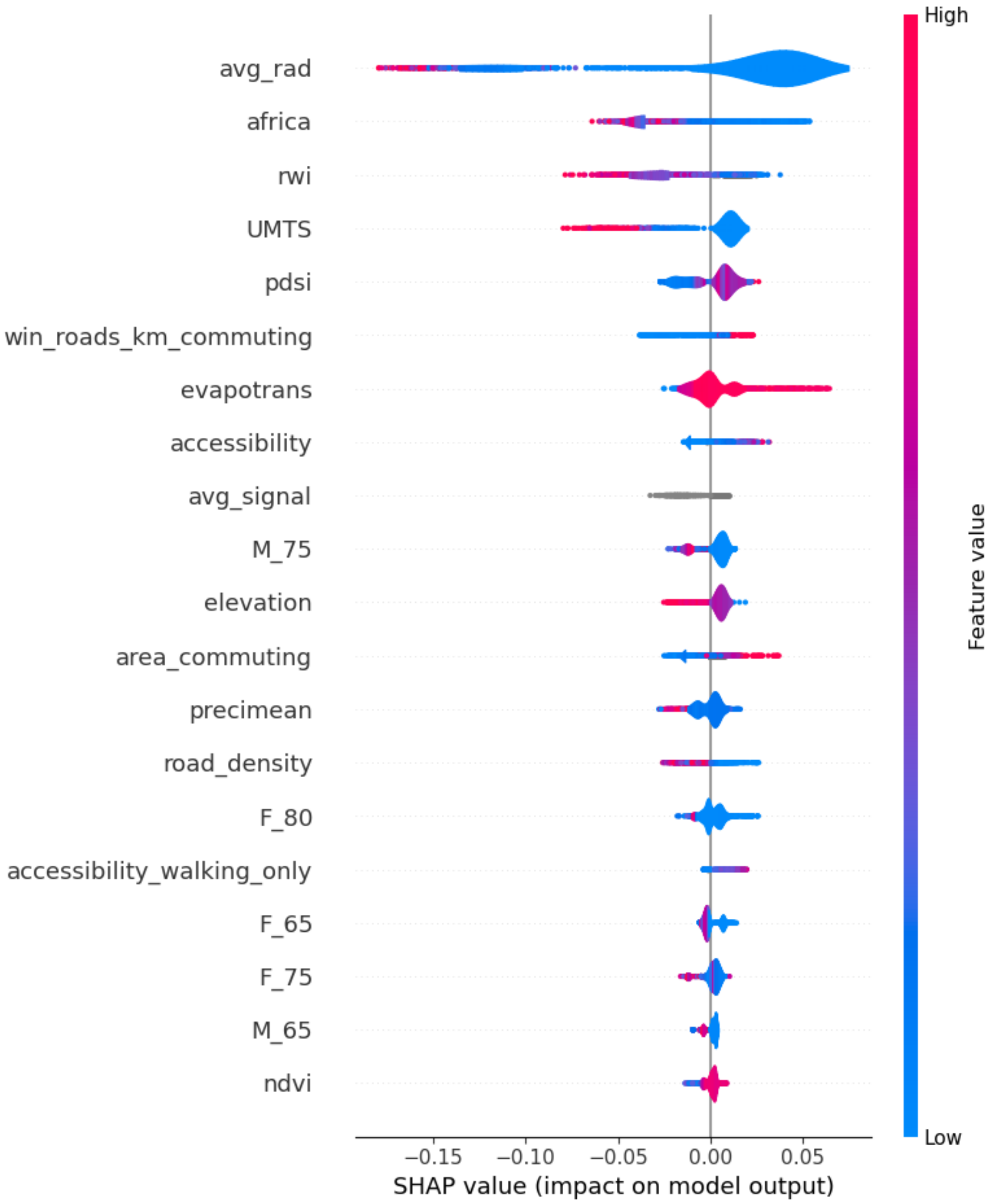


Figure 6.12: SHAP values for XGBoost model for prevalence. Here we can observe that the most important variable is `avg_rad`, that represents nighttime luminosity intensity, followed by `africa`, that represents the critical infrastructure spatial index. Gray represents NA values.

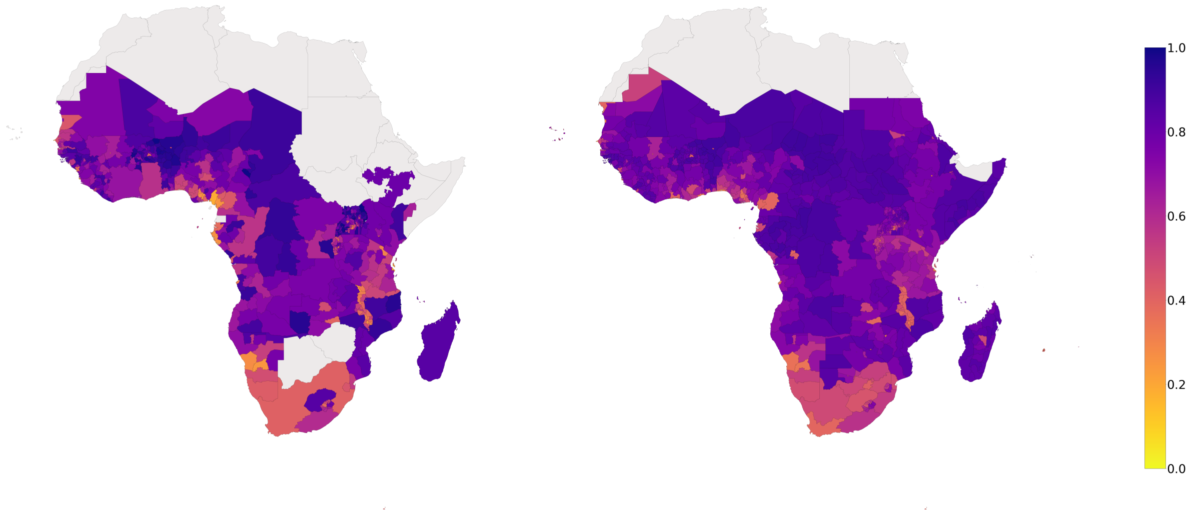


Figure 6.13: State level

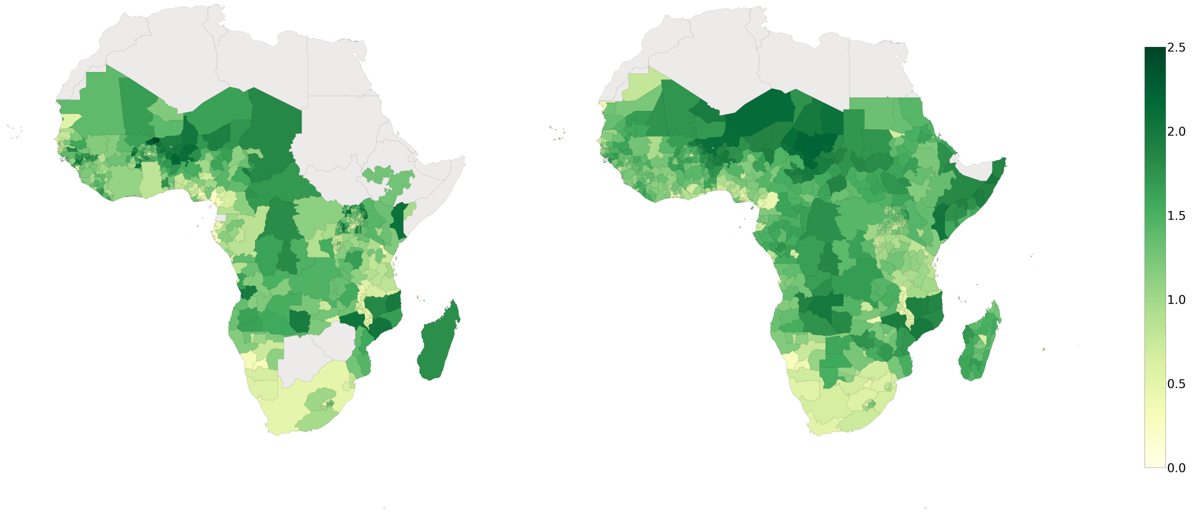


Figure 6.14: Caption

7 Conclusion

More than one billion children are multidimensionally poor. A child is multidimensionally poor if he/she is deprived in at least one dimension. The dimensions included in this internationally comparable multidimensional index are sanitation, water, housing, health, nutrition and nutrition. For each dimension we have just one indicator, that is shown along with the respective threshold and the age group considered in Table 1.1. We have equal weighting among the dimensions to construct the index, since all rights are equally important. In this study we focus on severe deprivations. Child poverty has negative long lasting effects on children, and its alarming consequences have made this the first point of the "Sustainable Development Goals" adopted in 2015 by the United Nations member states. To "end poverty in all its forms everywhere" it is essential to map the distribution of poverty. Studies in literature have contributed high resolution estimates mainly for asset-based poverty and consumption-based poverty. However, just a few focus on child poverty. It is essential to distinguish child poverty from adult poverty, and to define it independently from monetary poverty, since children have different rights from adults, they should not work and earn an income. Moreover, children constitute from 25% to 50% of the population, so measuring child poverty separately allows us to better track how poverty evolves. Therefore, child poverty is defined as the lack of resources needed to realize rights constitutive of poverty. These rights are the ones that depend on material resources for their realization.

Hence, the goal of the thesis is to construct a finely-grained map and to predict for each cell variables regarding poverty such as prevalence, depth, and the different deprivations. We build a hexagonal grid, delevoped by Uber [50], where each hexagon has on average an area of 5.16 km^2. The ground truth data derive from the DHS surveys, which have been processed to obtain a binary indicator for each dimension on whether the child is deprived in that dimension. The data have been aggregated at the hexagon level computing the mean. From there, we collect georeferenced data from several data sources such as Google Earth Engine, Uppsala Conflict Data Program, Open Street Map, Ookland Open Data, OpenCellID, Meta's Data for Good repository, WorldPop. A detailed description of the variables included can be found in Table 3.3. Hence, two experiments have been conducted.

1. In the first, we predict multidimensionally child poverty only in the 25 countries that have a recent DHS survey. The data have been processed with robust scaling and KNN imputer. A neighboring approach (explained in Section 3.1.2) has been implemented as a way to introduce smoothness, augment the data and account for DHS location displacement. For each country and for each dimension we compare the performance of XGBoost and LightGBM, finding the first slightly more performant in the majority of the cases. The R^2 results can be observed in Table 6.1 and in Table 6.2.

2. In the second, we build a model to be able to generalize on the countries that do not have DHS information. We randomly split the countries for which we have ground truth data in training and test, and we start modeling comparing spatial cross validation and the standard cross validation techniques. The two techniques perform similarly, with spatial cross validation slightly better. The results can be read in 6.3. Here we do not use the neighboring approach to have a more generalizable model, and we process the data with robust scaling and median imputer. The model implemented is XGBoost.

To facilitate responsible downstream use of the predictions, we include prediction intervals. The results for this model (point estimates and prediction intervals) for prevalence and depth can be seen in the Figures 6.8, 6.9, 6.10, 6.11. Moreover, we focus on interpreting the results using the SHAP values, since better interpretability leads to better adoption. The most important features can be observed in Figure 6.12. Lastly, aggregated predictions have been validated with the DHS and MICS subnational and national values, as can be observed in Figures 6.13, 6.14.

Comparing the results between the first and the second experiment, we observe that overall better results are achieved in the first one, and therefore the models fill the gaps of the DHS surveys. Hence, further DHS surveys in the remaining countries would improve the predictions in those countries.

Although localizing where poor children are is an important step towards the first Sustainable Goal, reducing poverty requires specific and targeted policies that vary from place to place and that need to be supported by local authorities and by the local community. This thesis does not cover the intricate and complex underlying causes of child poverty.

Including local knowledge and being ready to challenge our suppositions are crucial steps in overseas studies to achieve the goal of ending poverty.

Furthermore, while an uniform hexagonal grid has its advantages, it does not differentiate more populated places from less populated ones. Hence, this could be something to further investigate. Moreover, it is important to reflect how standardized procedures of reporting data may lead to systematic errors, excluding categories of children, such as street children. Moreover, a further point to investigate is differentiate the predictions on gender and age.

In conclusion, the goal of the thesis is to produce finely-grained poverty prediction of multidimensional child poverty. Different georefenced data are used to fill the missing values of DHS surveys, used as ground truth for the estimations.
The code can be found in `https://github.com/marinavicini/stc_continuing`

References

[1] UNICEF and Save the Children, "Impact of covid-19 on children living in poverty: A technical note," Tech. Rep., 2020.

[2] J. Currie and C. Spatz Widom, "Long-term consequences of child abuse and neglect on adult economic well-being," *Child maltreatment*, vol. 15, no. 2, pp. 111–120, 2010.

[3] OECD, *Changing the Odds for Vulnerable Children*, 2019. [Online]. Available: https://www.oecd-ilibrary.org/content/publication/a2e8796c-en

[4] United Nations General Assembly *et al.*, "Transforming our world: the 2030 agenda for sustainable development," *United Nations: New York, NY, USA*, 2015.

[5] UNICEF, "Child poverty profiles: Understanding internationally comparable estimates."

[6] ——, "Technical briefing note 1: Individual children and households," Tech. Rep.

[7] S. Kurukulasuriya and S. Engilbertsdóttir, "A multidimensional approach to measuring child poverty," *Child Poverty and New Inequality: New Perspectives*, pp. 23–34, 2012.

[8] M. Nowak and S. Osmani, "Human rights and poverty reduction - a conceptual framework," 2004.

[9] UNICEF, "Technical briefing note 2: the dimensions of child poverty," Tech. Rep.

[10] United Nations OHCHR, "Frequently asked questions on a human rights-based approach to development cooperation," *New York and Geneva*, 2006.

[11] UN General Assembly *et al.*, "Universal declaration of human rights," *UN General Assembly*, vol. 302, no. 2, pp. 14–25, 1948.

[12] UN Committee on the Rights of the Child, "General guidelines regarding the form and content of periodic reports to be submitted by states parties under article 44, paragraph 1 (b), of the convention," *UN Doc. CRC/C/58/Rev.1*, 2005.

[13] UNICEF, “Technical briefing note 3: Seven reasons for equal weighting of dimensions in child poverty measurement,” Tech. Rep.

[14] B. Eva, “Principles of indifference,” April 2019. [Online]. Available: http://philsci-archive.pitt.edu/16041/

[15] M. R. Hagerty and K. C. Land, “Constructing summary indices of quality of life: A model for the effect of heterogeneous importance weights,” *Sociological Methods & Research*, vol. 35, no. 4, pp. 455–496, 2007.

[16] D. Gordon, L. D. Howe, B. Galobardes, A. Matijasevich, D. Johnston, O. Onwujekwe, R. Patel, E. A. Webb, D. A. Lawlor, and J. R. Hargreaves, “Authors’ Response to: Alternatives to principal components analysis to derive asset-based indices to measure socio-economic position in low- and middle-income countries: the case for multiple correspondence analysis,” *International Journal of Epidemiology*, vol. 41, no. 4, pp. 1209–1210, 08 2012. [Online]. Available: https://doi.org/10.1093/ije/dys120

[17] C. Elbers, J. O. Lanjouw, and P. Lanjouw, “Micro-level estimation of poverty and inequality,” *Econometrica*, vol. 71, no. 1, pp. 355–364, 2003.

[18] B. S. Rowntree, *Poverty: A Study of Town Life*. Macmillan, 1902.

[19] T. Bedi, A. Coudouel, and K. Simler, *More than a pretty picture: using poverty maps to design better policies and interventions*. World Bank Publications, 2007.

[20] A. V. Banerjee, A. Deaton, N. Lustig, K. Rogoff, and E. Hsu, “An evaluation of world bank research, 1998-2005,” *Available at SSRN 2950327*, 2006.

[21] S. M. Xie, N. Jean, M. Burke, D. B. Lobell, and S. Ermon, “Transfer learning from deep features for remote sensing and poverty mapping,” in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, D. Schuurmans and M. P. Wellman, Eds. AAAI Press, 2016, pp. 3929–3935. [Online]. Available: http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12196

[22] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon, “Combining satellite imagery and machine learning to predict poverty,” *Science*, vol. 353, no. 6301, pp. 790–794, 2016.

[23] G. Chi, H. Fang, S. Chatterjee, and J. E. Blumenstock, “Microestimates of wealth for all low- and middle-income countries,” *Proceedings of the National Academy of Sciences*, vol. 119, no. 3, p. e2113658119, 2022. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.2113658119

[24] E. L. Aiken, G. Bedoya, J. Blumenstock, and A. Coville, “Program targeting with machine learning and mobile phone data: Evidence from an anti-poverty intervention in afghanistan,” *CoRR*, vol. abs/2206.11400, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2206.11400

[25] M. R. Khan and J. E. Blumenstock, “Multi-gcn: Graph convolutional networks for multi-view networks, with applications to global poverty,” in *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. AAAI Press, 2019, pp. 606–613. [Online]. Available: https://doi.org/10.1609/aaai.v33i01.3301606

[26] J. Blumenstock, G. Cadamuro, and R. On, “Predicting poverty and wealth from mobile phone metadata,” *Science*, vol. 350, no. 6264, pp. 1073–1076, 2015.

[27] J. E. Steele, P. R. Sundsøy, C. Pezzulo, V. A. Alegana, T. J. Bird, J. Blumenstock, J. Bjelland, K. Engø-Monsen, Y.-A. De Montjoye, A. M. Iqbal *et al.*, “Mapping poverty using mobile phone and satellite data,” *Journal of The Royal Society Interface*, vol. 14, no. 127, p. 20160690, 2017.

[28] P. O. Okwi, G. Ndeng’e, P. Kristjanson, M. Arunga, A. Notenbaert, A. Omolo, N. Henninger, T. Benson, P. Kariuki, and J. Owuor, “Spatial determinants of poverty in rural kenya,” *Proceedings of the National Academy of Sciences*, vol. 104, no. 43, pp. 16 769–16 774, 2007.

[29] K. Lee and J. Braithwaite, “High-resolution poverty maps in sub-saharan africa,” *CoRR*, vol. abs/2009.00544, 2020. [Online]. Available: https://arxiv.org/abs/2009.00544

[30] N. Puttanapong, A. Martinez, J. A. N. Bulan, M. Addawe, R. L. Durante, and M. Martillan, “Predicting poverty using geospatial data in thailand,” *ISPRS Int. J. Geo Inf.*, vol. 11, no. 5, p. 293, 2022. [Online]. Available: https://doi.org/10.3390/ijgi11050293

[31] D. Quercia, J. Ellis, L. Capra, and J. Crowcroft, "Tracking "gross community happiness" from tweets," in *CSCW '12 Computer Supported Cooperative Work, Seattle, WA, USA, February 11-15, 2012*, S. E. Poltrock, C. Simone, J. Grudin, G. Mark, and J. Riedl, Eds. ACM, 2012, pp. 965–968. [Online]. Available: https://doi.org/10.1145/2145204.2145347

[32] E. Sheehan, C. Meng, M. Tan, B. Uzkent, N. Jean, M. Burke, D. B. Lobell, and S. Ermon, "Predicting economic development using geolocated wikipedia articles," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, A. Teredesai, V. Kumar, Y. Li, R. Rosales, E. Terzi, and G. Karypis, Eds. ACM, 2019, pp. 2698–2706. [Online]. Available: https://doi.org/10.1145/3292500.3330784

[33] A. Perez, C. Yeh, G. Azzari, M. Burke, D. B. Lobell, and S. Ermon, "Poverty prediction with public landsat 7 satellite imagery and machine learning," *CoRR*, vol. abs/1711.03654, 2017. [Online]. Available: http://arxiv.org/abs/1711.03654

[34] C. Yeh, A. Perez, A. Driscoll, G. Azzari, Z. Tang, D. Lobell, S. Ermon, and M. Burke, "Using publicly available satellite imagery and deep learning to understand economic well-being in africa," *Nature communications*, vol. 11, no. 1, pp. 1–11, 2020.

[35] B. Babenko, J. Hersh, D. Newhouse, A. Ramakrishnan, and T. Swartz, "Poverty mapping using convolutional neural networks trained on high and medium resolution satellite images, with an application in mexico," *CoRR*, vol. abs/1711.06323, 2017. [Online]. Available: http://arxiv.org/abs/1711.06323

[36] J. Hersh, R. N. Engstrom, and M. L. Mann, "Open data for algorithms: mapping poverty in belize using open satellite derived features and machine learning," *Inf. Technol. Dev.*, vol. 27, no. 2, pp. 263–292, 2021. [Online]. Available: https://doi.org/10.1080/02681102.2020.1811945

[37] B. Tang, Y. Liu, and D. S. Matteson, "Predicting poverty with vegetation index," *Applied Economic Perspectives and Policy*, 2022.

[38] T. N. Croft, A. M. Marshall, C. K. Allen, F. Arnold, S. Assaf, S. Balian *et al.*, "Guide to dhs statistics," *Rockville: ICF*, vol. 645, 2018.

[39] M. E. Grosh and J. Muñoz, *A Manual for Planning and Implementing the Living Standards Measurement Study Survey*. The World Bank, 1996.

[40] NOAA National Geophysical Data Center, "F18 2013 nighttime lights composite," 2014.

[41] Microsoft, "Bing maps tile system - bing maps."

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 770–778. [Online]. Available: https://doi.org/10.1109/CVPR.2016.90

[43] J. Smits and R. Steendijk, "The international wealth index (iwi)," *Social indicators research*, vol. 122, no. 1, pp. 65–85, 2015.

[44] OpenStreetMap contributors, "Planet dump retrieved from https://planet.osm.org ," https://www.openstreetmap.org, 2017.

[45] M. Pesaresi, D. Ehrilch, A. J. Florczyk, S. Freire, A. Julea, T. Kemper, P. Soille, and V. Syrris, "Ghs built-up grid, derived from landsat, multitemporal (1975, 1990, 2000, 2014)," 2015.

[46] ——, "Ghs built-up confidence grid, derived from landsat, multitemporal (1975, 1990, 2000, 2014)," 2015.

[47] ——, "Hs built-up datamask grid derived from landsat, multitemporal (1975, 1990, 2000, 2014)," 2015.

[48] M. De Milliano and I. Plavgo, "Analysing multidimensional child poverty in sub-saharan africa: Findings using an international comparative approach," *Child indicators research*, vol. 11, no. 3, pp. 805–833, 2018.

[49] S. O. Rutstein, G. Rojas *et al.*, "Guide to dhs statistics," *Calverton, MD: ORC Macro*, vol. 38, p. 78, 2006.

[50] Uber Technologies Inc., "H3: Uber's hexagonal hierarchical spatial index." [Online]. Available: https://eng.uber.com/h3

[51] S. Davies, T. Pettersson, and M. Öberg, "Organized violence 1989–2021 and drone warfare," *Journal of Peace Research*, vol. 59, no. 4, pp. 593–610, 2022.

[52] R. Sundberg and E. Melander, "Introducing the ucdp georeferenced event dataset," *Journal of Peace Research*, vol. 50, no. 4, pp. 523–532, 2013.

[53] Wikimedia, "Commons: Mobile app." [Online]. Available: https://commons.wikimedia.org/wiki/Commons:Mobile_app

[54] L. Briem, M. Heilig, C. Klinkhardt, and P. Vortisch, "Analyzing openstreetmap as data source for travel demand models a case study in karlsruhe," *Transportation Research Procedia*, vol. 41, pp. 104–112, 2019, urban Mobility – Shaping the Future Together mobil.TUM 2018 – International Scientific Conference on Mobility and Transport Conference Proceedings. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352146519304338

[55] Unwired Labs, "Opencellid." [Online]. Available: https://opencellid.org/

[56] S. Nirandjan, E. E. Koks, P. J. Ward, and J. C. Aerts, "A spatially-explicit harmonized global dataset of critical infrastructure," *Scientific Data*, vol. 9, no. 1, pp. 1–13, 2022.

[57] M. Kummu, M. Taka, and J. H. Guillaume, "Gridded global datasets for gross domestic product and human development index over 1990–2015," *Scientific data*, vol. 5, no. 1, pp. 1–15, 2018.

[58] Data for Good at Meta, "Commuting zones." [Online]. Available: https://dataforgood.facebook.com/dfg/tools/commuting-zones

[59] A. Jarvis, H. I. Reuter, A. Nelson, E. Guevara *et al.*, "Hole-filled srtm for the globe version 4," *available from the CGIAR-CSI SRTM 90m Database (http://srtm. csi. cgiar. org)*, vol. 15, no. 25-54, p. 5, 2008.

[60] J. T. Abatzoglou, S. Z. Dobrowski, S. A. Parks, and K. C. Hegewisch, "Terraclimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015," *Scientific data*, vol. 5, no. 1, pp. 1–12, 2018.

[61] G. Huffman, E. Stocker, D. Bolvin, E. Nelkin, and T. Jackson, "Pm imerg final precipitation l3 1 month 0.1 degree x 0.1 degree v06," 2019.

[62] D. Weiss, A. Nelson, C. Vargas-Ruiz, K. Gligorić, S. Bavadekar, E. Gabrilovich, A. Bertozzi-Villa, J. Rozier, H. Gibson, T. Shekel *et al.*, "Global maps of travel time to healthcare facilities," *Nature Medicine*, vol. 26, no. 12, pp. 1835–1838, 2020.

[63] A. Lyapustin and Y. Wang, "Mcd19a2 modis/terra+ aqua land aerosol optical depth daily l2g global 1km sin grid v006 [data set]," *NASA EOSDIS land processes DAAC*, 2018.

[64] A. J. Tatem, "Worldpop, open data for spatial demography," *Scientific Data*, vol. 4, no. 1, Jan 2017. [Online]. Available: http://dx.doi.org/10.1038/sdata.2017.4

[65] C. Linard, M. Gilbert, R. W. Snow, A. M. Noor, and A. J. Tatem, "Population distribution, settlement patterns and accessibility across africa in 2010," *PLOS ONE*, vol. 7, no. 2, pp. 1–8, 02 2012. [Online]. Available: https://doi.org/10.1371/journal.pone.0031743

[66] O. G. Troyanskaya, M. N. Cantor, G. Sherlock, P. O. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for DNA microarrays," *Bioinform.*, vol. 17, no. 6, pp. 520–525, 2001. [Online]. Available: https://doi.org/10.1093/bioinformatics/17.6.520

[67] G. Shmueli, "To explain or to predict?" *Statistical science*, vol. 25, no. 3, pp. 289–310, 2010.

[68] V. Vapnik, "An overview of statistical learning theory," *IEEE Trans. Neural Networks*, vol. 10, no. 5, pp. 988–999, 1999. [Online]. Available: https://doi.org/10.1109/72.788640

[69] ——, "Principles of risk minimization for learning theory," pp. 831–838, 1991. [Online]. Available: http://papers.nips.cc/paper/506-principles-of-risk-minimization-for-learning-theory

[70] A. Tewari and P. L. Bartlett, "Learning theory," in *Academic Press Library in Signal Processing*. Elsevier, 2014, vol. 1, pp. 775–816.

[71] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Wadsworth, 1984.

[72] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[73] S. Milborrow, "Titanic decision tree." 2011. [Online]. Available: https://commons.wikimedia.org/wiki/File:CART_tree_titanic_survivors.png

[74] H. Laurent and R. L. Rivest, "Constructing optimal binary decision trees is np-complete," *Information processing letters*, vol. 5, no. 1, pp. 15–17, 1976.

[75] R. E. Schapire, "The strength of weak learnability," *Mach. Learn.*, vol. 5, pp. 197–227, 1990. [Online]. Available: https://doi.org/10.1007/BF00116037

[76] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *Machine Learning, Proceedings of the Thirteenth International Conference (ICML '96), Bari, Italy, July 3-6, 1996*, L. Saitta, Ed. Morgan Kaufmann, 1996, pp. 148–156.

[77] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[78] ——, "Stochastic gradient boosting," *Computational statistics & data analysis*, vol. 38, no. 4, pp. 367–378, 2002.

[79] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996. [Online]. Available: https://doi.org/10.1007/BF00058655

[80] D. W. Opitz and J. W. Shavlik, "Generating accurate and diverse members of a neural-network ensemble," pp. 535–541, 1995. [Online]. Available: http://papers.nips.cc/paper/1175-generating-accurate-and-diverse-members-of-a-neural-network-ensemble

[81] ——, "Actively searching for an effective neural network ensemble," *Connect. Sci.*, vol. 8, no. 3, pp. 337–354, 1996. [Online]. Available: https://doi.org/10.1080/095400996116802

[82] M. Sheykhmousa, M. MahdianPari, H. Ghanbari, F. Mohammadimanesh, P. Ghamisi, and S. Homayouni, "Support vector machine versus random forest for remote sensing image classification: A meta-analysis and systematic review," *IEEE J. Sel. Top.*

Appl. Earth Obs. Remote. Sens., vol. 13, pp. 6308–6325, 2020. [Online]. Available: https://doi.org/10.1109/JSTARS.2020.3026724

[83] A. Natekin and A. C. Knoll, "Gradient boosting machines, a tutorial," *Frontiers Neurorobotics*, vol. 7, p. 21, 2013. [Online]. Available: https://doi.org/10.3389/fnbot.2013.00021

[84] M. Ll and J. Baxter, "Boosting algorithms as gradient descent in function space," 1999.

[85] L. Mason, J. Baxter, P. L. Bartlett, and M. R. Frean, "Boosting algorithms as gradient descent," pp. 512–518, 1999. [Online]. Available: http://papers.nips.cc/paper/1766-boosting-algorithms-as-gradient-descent

[86] T. G. Dietterich, A. Ashenfelter, and Y. Bulatov, "Training conditional random fields via gradient tree boosting," in *Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004), Banff, Alberta, Canada, July 4-8, 2004*, ser. ACM International Conference Proceeding Series, C. E. Brodley, Ed., vol. 69. ACM, 2004. [Online]. Available: https://doi.org/10.1145/1015330.1015428

[87] P. A. Viola and M. J. Jones, "Rapid object detection using a boosted cascade of simple features," in *2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001), with CD-ROM, 8-14 December 2001, Kauai, HI, USA*. IEEE Computer Society, 2001, pp. 511–518. [Online]. Available: https://doi.org/10.1109/CVPR.2001.990517

[88] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, B. Krishnapuram, M. Shah, A. J. Smola, C. C. Aggarwal, D. Shen, and R. Rastogi, Eds. ACM, 2016, pp. 785–794. [Online]. Available: https://doi.org/10.1145/2939672.2939785

[89] Microsoft, "Lightgbm's documentation." [Online]. Available: https://lightgbm.readthedocs.io/en/v3.3.2/#

[90] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," pp. 3146–3154, 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

[91] H. Shi, "Best-first decision tree learning," Ph.D. dissertation, The University of Waikato, 2007.

[92] T. R. Jensen and B. Toft, *Graph Coloring Problems*. John Wiley & Sons, 2011.

[93] W. R. Tobler, "A computer movie simulating urban growth in the detroit region," *Economic geography*, vol. 46, no. sup1, pp. 234–240, 1970.

[94] S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions," pp. 4765–4774, 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[95] V. Taquet, V. Blot, T. Morzadec, L. Lacombe, and N. Brunel, "MAPIE: an open-source library for distribution-free uncertainty quantification," *CoRR*, vol. abs/2207.12274, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2207.12274

[96] Y. Romano, E. Patterson, and E. J. Candès, "Conformalized quantile regression," pp. 3538–3548, 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/5103c3584b063c431bd1268e9b5e76fb-Abstract.html

[97] M. Sadinle, J. Lei, and L. A. Wasserman, "Least ambiguous set-valued classifiers with bounded error levels," *CoRR*, vol. abs/1609.00451, 2016. [Online]. Available: http://arxiv.org/abs/1609.00451

[98] A. N. Angelopoulos, S. Bates, M. I. Jordan, and J. Malik, "Uncertainty sets for image classifiers using conformal prediction," 2021. [Online]. Available: https://openreview.net/forum?id=eNdiU_DbM9

[99] B. Kim, C. Xu, and R. F. Barber, "Predictive inference is free with the jackknife+-after-bootstrap," 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/2b346a0aa375a07f5a90a344a61416c4-Abstract.html

[100] Y. Romano, M. Sesia, and E. J. Candès, "Classification with valid and adaptive coverage," 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/244edd7e85dc81602b7615cd705545f5-Abstract.html

[101] R. F. Barber, E. J. Candes, A. Ramdas, and R. J. Tibshirani, "Predictive inference with the jackknife+," *The Annals of Statistics*, vol. 49, no. 1, pp. 486–507, 2021.

[102] C. Xu and Y. Xie, "Conformal prediction interval for dynamic time-series," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 2021, pp. 11 559–11 569. [Online]. Available: http://proceedings.mlr.press/v139/xu21h.html

[103] C. Wang, Q. Wu, M. Weimer, and E. Zhu, "FLAML: A fast and lightweight automl library," in *Proceedings of Machine Learning and Systems 2021, MLSys 2021, virtual, April 5-9, 2021*, A. Smola, A. Dimakis, and I. Stoica, Eds. mlsys.org, 2021. [Online]. Available: https://proceedings.mlsys.org/paper/2021/hash/92cc227532d17e56e07902b254dfad10-Abstract.html

Acknowledgments

This thesis is the continuation of a project started during the summer of 2022 at the Data Science for Social Good (DSSG) program at the University of Warwick, in collaboration with Save the Children and UNICEF. During the fellowship I was part of a team composed by Daniela Pinto Veizaga, John Fitzgerald and Arpita Saggar. Together we implemented a methodology to obtain child poverty predictions from different georeferenced data sources, as a proof of concept. After the summer, I continued the project with a consultancy position at Save the Children, expanding it to all the remaining sub-Saharan African countries.

This project has been presented at the "Learning and Doing Data for Good" conference held in Seattle in September 2022. The presentation can be found here: `https://youtu.be/R18xqwdpB0s?t=5883`.

Thanks to Lamberto Ballan, my thesis supervisor, for trusting me during the process.
Thanks to Juergen Branke, director of DSSGx Warwick, for all your efforts in organizing this special program.
Thanks to my team at DSSG for all your precious work, from which this thesis have been built on and from all that I have learned from you.
Thanks to Enrique Delamonica, UNICEF Statistics and Monitoring Senior Adviser for Child Poverty and Gender Equality, for your valuable input and feedback.
Thanks to Oliver Fiala, Senior Research Adviser at Save the Children, for your continuous collaboration. It has been a real pleasure working together.