

# Strategische Übertragbarkeit bei der Inferenz von Armutsverteilungskarten

DIPLOMARBEIT

zur Erlangung des akademischen Grades

**Diplom-Ingenieur**

im Rahmen des Studiums

**Data Science**

eingereicht von

**Yannik Gaebel**

Matrikelnummer 12208157

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Univ. Prof. Dr. Allan Hanbury

Mitwirkung: Dr. Lisette Espín-Noboa

Prof. Dr. János Kertész

Prof. Dr. Márton Karsai

Wien, 3. Dezember 2025

---

Yannik Gaebel

---

Allan Hanbury



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Strategic Transferability in Poverty Map Inference

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

**Diplom-Ingenieur**

in

**Data Science**

by

**Yannik Gaebel**

Registration Number 12208157

to the Faculty of Informatics

at the TU Wien

Advisor: Univ. Prof. Dr. Allan Hanbury

Assistance: Dr. Lisette Espín-Noboa

Prof. Dr. János Kertész

Prof. Dr. Márton Karsai

Vienna, December 3, 2025

---

Yannik Gaebel

---

Allan Hanbury



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Erklärung zur Verfassung der Arbeit

Yannik Gaebel

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Ich erkläre weiters, dass ich mich generativer KI-Tools lediglich als Hilfsmittel bedient habe und in der vorliegenden Arbeit mein gestalterischer Einfluss überwiegt. Im Anhang „Übersicht verwendeter Hilfsmittel“ habe ich alle generativen KI-Tools gelistet, die verwendet wurden, und angegeben, wo und wie sie verwendet wurden. Für Textpassagen, die ohne substantielle Änderungen übernommen wurden, haben ich jeweils die von mir formulierten Eingaben (Prompts) und die verwendete IT- Anwendung mit ihrem Produktnamen und Versionsnummer/Datum angegeben.

Wien, 3. Dezember 2025

---

Yannik Gaebel



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Kurzfassung

Maschinelles Lernen bietet eine vielversprechende Lösung für die Erstellung hochauflösender Armutskarten, doch wird seine Anwendung in datenarmen Regionen wie Subsahara-Afrika häufig durch begrenzte verfügbare Erhebungsdaten und unvollständige Geodaten erschwert. Diese Arbeit befasst sich mit diesen Herausforderungen, indem Strategien untersucht werden, um die Übertragbarkeit von Modellen zwischen Ländern zu verbessern und den Umgang mit fehlenden Daten für eine präzise Armutsschätzung zu optimieren. Anhand von Daten aus sechs unterschiedlichen subsaharischen Ländern und vier Datenquellen (Nachlichter, Bevölkerung, Mobilfunk und Infrastruktur) werden CatBoost-Modelle eingesetzt, um die Rolle der Ländersimilarität für den Modelltransfer zu evaluieren, optimale Strategien zum Umgang mit fehlenden Daten zu bestimmen und die Wirksamkeit verschiedener Transfer-Learning-Techniken zu vergleichen.

Die Ergebnisse zeigen, dass die Transferleistung stark von der Ländersimilarität abhängt, wobei der Jones Country Similarity Index besonders aussagekräftig ist. Zudem spielt die Auswahl der Länder, die für das Modelltraining genutzt werden, eine zentrale Rolle: Die Hinzunahme ähnlicher Länder kann die Leistung verbessern, während die Einbeziehung unähnlicher häufig zu negativem Transfer führt. Die Modelle zeigten eine hohe Robustheit gegenüber teilweisem Datenverlust, während sich die Rekonstruktion fehlender Featurekategorien als weniger wirksam erweist. Unter den Transfer-Learning-Methoden zeigte Feature Augmentation die höchste Wirksamkeit und übertraf in fünf von sechs Ländern die Baseline-Ergebnisse. Die Ergebnisse sind in einem praxisnahen Entscheidungsrahmenwerk für die Armutskartierung in datenarmen Umgebungen zusammengefasst.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Abstract

Machine learning offers a promising solution for high-resolution poverty mapping, but its application in data-scarce regions like sub-Saharan Africa is often hampered by limited ground-truth survey data and incomplete geospatial features. This thesis addresses these challenges by investigating strategies to enhance model transferability and manage missing data for accurate poverty estimation. Using data from six diverse sub-Saharan African countries and four feature sources (nighttime lights, population, cell towers, and infrastructure), this study employs CatBoost models to evaluate the role of country similarity in model transfer, determine optimal strategies for handling missing data, and benchmark the effectiveness of various transfer learning techniques.

The findings show that transfer performance is strongly influenced by country similarity, with the Jones Country Similarity Index proving most predictive. Moreover, the choice of source countries matters greatly: adding data from similar countries can improve performance, whereas including dissimilar ones often introduces negative transfer. Re-training on available features consistently outperformed reconstructing missing ones, indicating model robustness to partial data loss. Among transfer learning methods, Feature Augmentation was most effective, outperforming within-country baselines in five of six countries. The study contributes a practical framework for poverty mapping in data-scarce environments.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Contents

<b>Kurzfassung</b>	<b>vii</b>
<b>Abstract</b>	<b>ix</b>
<b>Contents</b>	<b>xi</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 Literature Review</b>	<b>7</b>
2.1 Poverty Mapping . . . . .	7
2.2 Transfer Learning for Tabular Data . . . . .	11
2.3 Country Similarity Indices . . . . .	13
<b>3 Data</b>	<b>17</b>
3.1 Selection of Six Sub-Saharan African Countries for Analysis . . . . .	17
3.2 Ground-truth Data . . . . .	17
3.3 Feature Sources . . . . .	19
<b>4 Methods</b>	<b>21</b>
4.1 Machine Learning Architectures . . . . .	21
4.2 Experimental Setup . . . . .	27
4.3 Country Similarity Indices . . . . .	29
<b>5 Results</b>	<b>35</b>
5.1 Poverty Map Models . . . . .	35
5.2 Model Transfer . . . . .	37
5.3 Transfer Learning . . . . .	44
5.4 Transferability decision framework . . . . .	47
<b>6 Conclusion</b>	<b>49</b>
6.1 Discussion and Contributions . . . . .	49
6.2 Limitations and Future Research . . . . .	51
<b>Übersicht verwendeter Hilfsmittel</b>	<b>53</b>
	xi

<b>List of Figures</b>	<b>55</b>
<b>List of Tables</b>	<b>56</b>
<b>Bibliography</b>	<b>57</b>

# CHAPTER 1

## Introduction

Ending poverty in all its forms is the first United Nations Sustainable Development Goal. Yet, extreme poverty remains a persistent challenge, particularly in sub-Saharan Africa, where it undermines health outcomes, economic development, social cohesion, and climate resilience [HS02]. Poverty maps serve as a valuable tool to represent the geographic distribution of wealth, providing insights for policymakers, NGOs, and researchers. These maps inform targeted interventions and policy decisions [YPD<sup>+</sup>20].

However, traditional data sources for poverty mapping, such as census and household surveys, are often outdated, expensive and challenging to obtain, especially in conflict zones or during pandemics. Consequently, machine learning models have emerged as a promising approach for poverty mapping, leveraging their ability to predict wealth distribution with high spatial resolution. By training on diverse data, including satellite imagery and metadata from mobile and web platforms, these models enable rapid and accurate inference, addressing the limitations of conventional data sources.

A prominent method involves using satellite imagery to generate poverty maps [MD24]. Recent studies demonstrate that combining daytime and nighttime light satellite images enhances predictive accuracy [YPD<sup>+</sup>20]. These models have shown robust performance, with the ability to generalize across countries and scale effectively to produce nationwide poverty maps.

In addition to satellite imagery, other data sources for poverty prediction include mobile phone usage records, connectivity infrastructure, indicators of economic activity, accessibility to services, population estimates, and physical geography [PJ17, PWP23]. Studies demonstrate that combining multiple data sources typically yields superior performance [PWP23, PJ17]. Satellite-based estimates, for instance, have been shown to perform better at distinguishing wealthy areas from poor ones but are less effective in separating poor from near-poor communities [YPD<sup>+</sup>20]. This limitation underscores the value of incorporating structured metadata such as connectivity indicators or mobile phone records,

which have been found particularly useful in rural or less urbanized areas [KKEN24]. The strengths of one modality may compensate for the weaknesses of another, improving the model's ability to generalize across heterogeneous settings.

The literature demonstrates that machine learning-based poverty maps can serve as practical tools to inform targeted social interventions. For instance, social protection programs often rely on asset-based targeting, where villages below a specific asset threshold receive aid, while those above do not. Such approaches, traditionally based on survey-derived asset data, benefit significantly from machine learning-driven poverty maps [YPD<sup>+</sup>20]. Despite these advances, several limitations persist. Fairness concerns arise due to systematic biases in satellite-based models, which often rely on urbanization signals and may misrepresent wealth rankings between urban and rural areas [ARB23]. Additionally, capturing temporal changes in wealth remains challenging, with models typically exhibiting modest performance in monitoring poverty dynamics over time.

Another critical challenge is the assumption that ground-truth labels and all relevant features are available for training and testing models. In practice, data availability is often limited, particularly in regions with low technology adoption, privacy restrictions, or other local constraints. For example, between 2000 and 2010, 39 out of 59 African countries conducted fewer than two nationally representative surveys suitable for poverty estimation, and 14 conducted none [JBX<sup>+</sup>16]. Even when surveys are conducted, much of the data remains inaccessible to researchers and policymakers. These data gaps hinder the ability to monitor poverty trends and design effective interventions, particularly in the most vulnerable regions.

This thesis addresses the critical challenges posed by limited data availability in machine learning-based poverty mapping, with a specific focus on enhancing model transferability and managing missing data. A key obstacle in applying machine learning models to poverty mapping in data-scarce regions is the lack of effective model transfer. When ground-truth poverty data is unavailable for a target country, models trained on data from other source countries can be leveraged. Since no adaptation to the local context is possible, model transfer relies on the generalizability of the learned patterns from the source domains. The success of such transfer therefore depends on selecting a source country with data characteristics and poverty correlates closely aligned with those of the target country. To facilitate optimal source country selection, various country similarity metrics can be employed. These metrics quantify relatedness based on factors such as geographic proximity, socio-economic indicators, environmental conditions, or distributions of remotely sensed features. The core hypothesis is that a model transferred from a source country deemed highly similar to the target country by a robust similarity metric will achieve higher predictive accuracy compared to a model from a less similar source.

Transfer learning is a machine learning approach in which a model trained on one dataset or task is adapted to a related dataset or task. In the context of poverty mapping, transfer learning involves fine-tuning a model from a source country using the typically scarce ground-truth data available in the target country. The aim is to enhance

---

predictive accuracy in the target country by transferring knowledge from the source country, utilizing patterns learned during initial model training. While widely used in other machine learning domains, its application to poverty mapping remains relatively underexplored.

Data from six economically diverse sub-Saharan African countries are used. Four feature sources, obtained at no cost, are integrated: nighttime light-based features from Google Earth Engine, population density data from Meta’s Data for Good initiative, cell tower GPS positions from OpenCellID for connectivity features, and infrastructure data from OpenStreetMap. The CatBoost algorithm is used to build models with a ground-truth wealth index derived from survey data to estimate socioeconomic patterns. To create the poverty map, the model is subsequently applied to all populated places. For illustration, the resulting inferred poverty maps for Sierra Leone and Uganda are shown in Figure 1.1 [ENKK23].

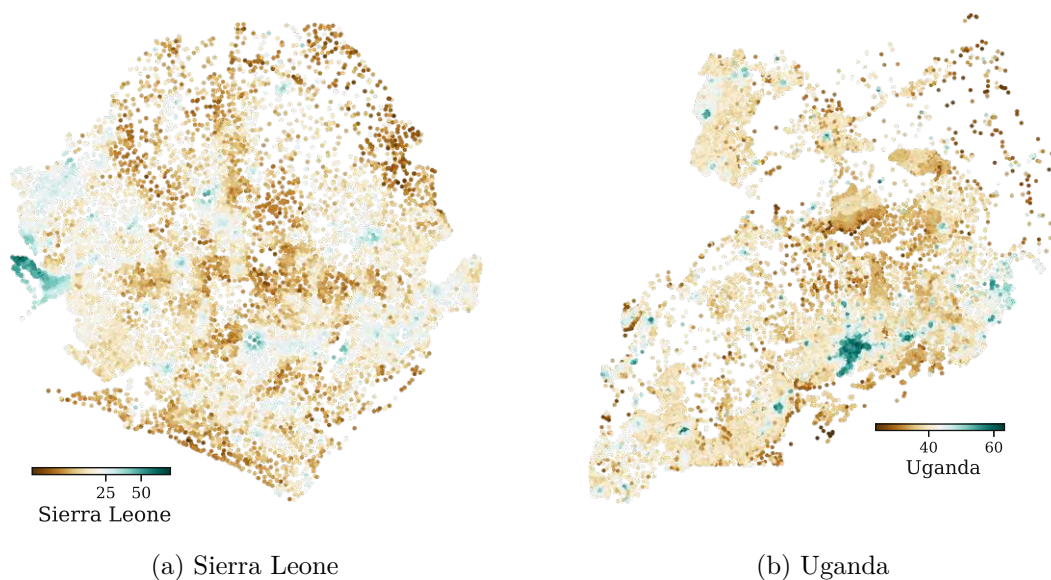


Figure 1.1: Inferred poverty maps for Sierra Leone (left) and Uganda (right), with color scales representing estimated values of the International Wealth Index [ENKK23].

Each survey cluster typically comprises 25–30 households. While poverty maps usually predict only the cluster mean, we train CatBoost models to estimate both the cluster mean and standard deviation, capturing a richer picture of local socioeconomic conditions. The standard deviation adds another dimension for understanding vulnerability and inequality, offering more detailed insights for designing targeted interventions.

The methodology for generating poverty maps in this study is illustrated in Figure 1.2.

As a next step, the transferability of the developed CatBoost models to new target countries is evaluated. This includes comparing single-source and multi-source transfer

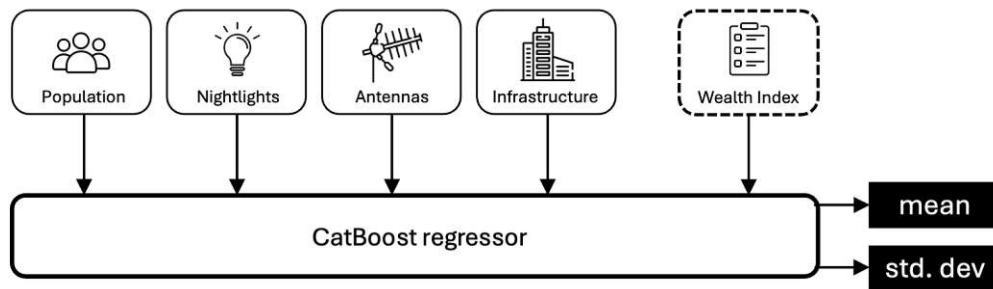


Figure 1.2: Overview of the fundamental machine learning setup used to obtain poverty maps. After Espin-Noboa et al. (2023) [ENKK23].

strategies and assessing how country similarity, measured via structural and feature-based metrics, informs optimal model selection. The experimental setup for evaluating model transferability is depicted in Figure 1.3.

In addition, scenarios with incomplete data are simulated to compare strategies for handling missing feature categories and benchmark several transfer learning techniques, including Multi-country training, Model Stacking, TransTab, CORAL, TrAdaBoostR2, and Feature Augmentation. Multi-country training pools data from all countries while controlling for systematic differences with a country indicator. Model Stacking combines predictions from single-source models as additional inputs for the target model. TransTab is a deep learning approach designed for transfer on tabular data. CORAL aligns the covariance structure of source and target features to reduce distributional differences. TrAdaBoostR2 iteratively reweights samples to emphasize informative target data. Finally, Feature Augmentation explicitly separates shared and domain-specific features, enabling the model to leverage commonalities while preserving country-specific patterns. Together, these analyses aim to disentangle the conditions under which model transfer is effective and provide a framework for data-efficient poverty mapping in low-resource settings.

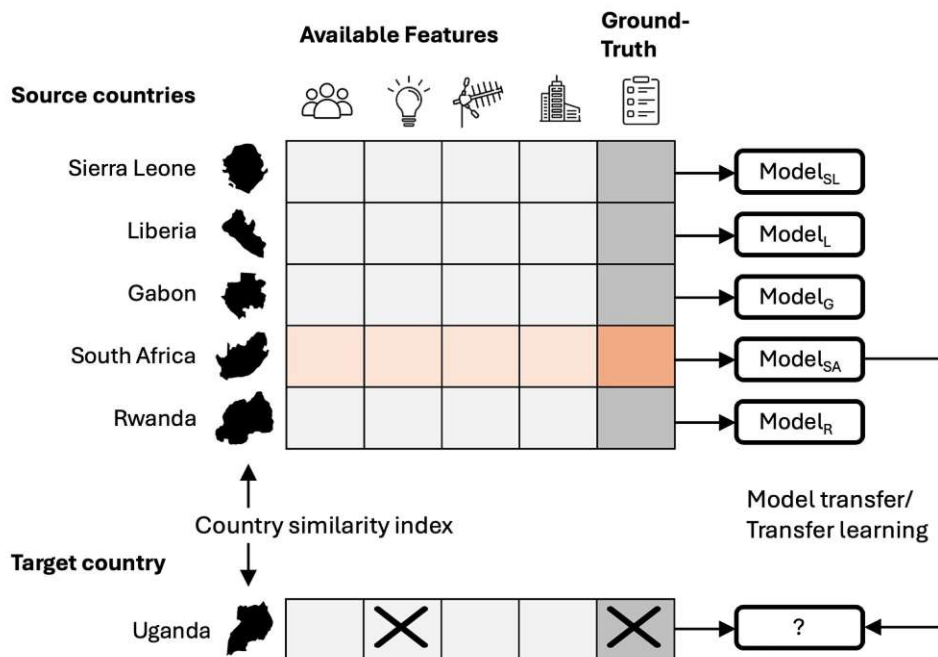


Figure 1.3: This figure presents a scenario where a poverty map needs to be created for Uganda (target country), despite lacking a complete feature layer and ground-truth data. In this case, model transfer offers a solution by leveraging pre-trained models from similar countries, identified through a country similarity index. The most similar country, highlighted in orange, represents a suitable source for model transfer. If ground-truth data were available in Uganda, transfer learning could be used by loading the model from the source country and continuing training with data from Uganda.

The research questions are as follows:

- RQ1** By how much does country similarity affect the performance of model transfer? If so, what is the most effective similarity metric for identifying the best country from which a model can be re-trained or transferred to produce a poverty map in another country?
- RQ2** What is the optimal strategy for generating a target country's poverty map when ground truth and some features are missing in a target country, but models and data from other countries are available? Should we (a) re-train models from other countries using the available features, or (b) reconstruct the missing features in the country before applying model transfer?
- RQ3** To what extent can transfer learning enhance predictive performance over models trained solely on in-country data?



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Literature Review

This chapter reviews the existing literature relevant to poverty mapping and transfer learning for tabular data. It begins by exploring the evolution of poverty mapping, from traditional census-based approaches to modern methods leveraging diverse data sources such as satellite imagery and mobile metadata. The chapter also examines cross-country training and model transferability, emphasizing the role of country similarity. Additionally, it covers transfer learning frameworks for tabular data, including gradient-boosted decision trees and deep tabular models. Finally, it presents the concept of country similarity indices and their application in improving transfer learning outcomes. Together, these sections establish foundations for the thesis, identifying gaps and opportunities in data-efficient poverty mapping for low-resource settings.

## 2.1 Poverty Mapping

Accurate and detailed poverty maps are important tools for targeting interventions in Sub-Saharan Africa. These maps offer snapshots of territorial poverty distributions, enabling informed decision-making for policymakers and stakeholders. They serve multiple practical purposes in governance and development planning. At the national level, poverty maps assist in resource allocation among provinces and districts, facilitating more equitable distribution of development funds. At more localized scales, they guide the implementation of targeted poverty alleviation initiatives within cities and towns. Poverty mapping has been widely utilized to guide social, agricultural, emergency response, environmental management, and broader anti-poverty programs [SB22] [MTB13].

Traditionally, poverty mapping has relied on a combination of census and household survey data. A key early approach was small area poverty estimation, which combines two data sources: a detailed household survey and a population census. The household survey is used to estimate a statistical model of welfare, which is then applied to the census data to predict poverty indicators in areas not covered by the survey [ELL03].

However, traditional methods have significant limitations, including substantial temporal gaps, high costs, lengthy data collection periods, and inadequate coverage of remote or inaccessible areas. Consequently, poverty mapping methodologies have progressively evolved to incorporate diverse and complementary data sources.

### 2.1.1 Machine Learning for Poverty Mapping

Recent advances in poverty mapping have increasingly relied on machine learning techniques to overcome limitations inherent in traditional methods, especially the scarcity of high-resolution socio-economic data. Researchers have integrated diverse sources such as satellite imagery, nighttime luminosity, mobile phone metadata, crowd-sourced information, and infrastructure indicators to improve predictive accuracy and coverage [SFM<sup>+</sup>18]. Prominent examples using satellite imagery include studies by Jean et al. (2016) and Yeh et al. (2020).

Jean et al. (2016) presented a transfer learning pipeline that predicts local poverty using satellite imagery [JBX<sup>+</sup>16]. The approach begins with a CNN pretrained on ImageNet to leverage general visual features. This model is then fine-tuned to predict nighttime light intensity from daytime satellite images, using large-scale, globally available data. This proxy task helps the network learn socioeconomic visual cues (e.g., roads, rooftops) without requiring household survey data. Finally, fixed CNN features are averaged at the survey-cluster level, reduced via PCA, and mapped to economic outcomes using ridge regression. The use of nightlights as an intermediate task circumvents the scarcity of labeled poverty data and significantly boosts performance, especially in rural regions where raw nightlight data underperform. Models trained across multiple countries approach the accuracy of country-specific models.

Extending the framework of Jean et al. (2016), Kondmann et al. (2020) assess the temporal robustness of models for poverty prediction, using Rwanda as a case study [KZ20]. They apply a ResNet-based model to multi-year Landsat 7 imagery to estimate changes in local wealth from 2005 to 2015. Although the model successfully captures spatial wealth patterns, it fails to reflect actual temporal dynamics. The study highlights a key limitation of current satellite-based models: while effective for static poverty mapping, they struggle to detect economic changes over time, likely due to insufficient temporal sensitivity in low-resolution satellite imagery and feature representations.

Building on this work, Yeh et al. (2020) developed an end-to-end deep learning model that directly incorporates both daytime multispectral Landsat imagery and nighttime luminosity data to estimate asset wealth across approximately 20,000 African villages [YPD<sup>+</sup>20]. Unlike earlier approaches, their joint training procedure allows the network to directly identify and leverage wealth-related visual patterns. This method substantially outperforms previous approaches, explaining up to 70% of the variation in wealth across multiple held-out countries. Importantly, the authors demonstrated the model's value in capturing established climate-wealth relationships and effectively targeting hypothetical social protection programs, thereby confirming its utility for research and

policy. Nonetheless, they emphasized the inherent trade-off between predictive accuracy and interpretability, advocating for the complementary use of satellite-based predictions alongside traditional household surveys.

The practical implications of employing satellite-derived poverty estimates for policy interventions, particularly regarding fairness, were further explored by Aiken et al. (2023) [ARB23]. They evaluate how prediction errors and representation disparities affect the use of satellite-derived wealth estimates for policy. Using satellite data from ten countries, they show that models are good at distinguishing between urban and rural areas but perform less effectively within either category, suggesting models rely strongly on urbanization signals rather than finer-grained indicators of poverty. They also find systematic biases in the rank ordering of wealth: in some countries, rural areas are consistently over-ranked, while in others, they are under-ranked relative to urban areas. These biases lead to allocative disparities in simulated aid programs, with poorer regions often misidentified, especially in contexts where urbanization and poverty are weakly correlated.

These findings highlight crucial considerations for policymakers, underscoring the necessity for high model precision and attention to representational fairness when deploying such estimates for targeting social programs. To enhance predictive performance and mitigate the limitations of single-source models, recent literature emphasizes the integration of multiple feature sources [LB22]. Each data source brings unique advantages: image-based features are particularly beneficial in urban environments, while structured metadata such as mobile connectivity and phone usage patterns offer stronger predictive signals in rural areas [KKEN24]. Examples demonstrating the advantages of multi-source approaches include the studies by Pokhriyal and Jacques (2017) and Putri et al. (2023).

Pokhriyal and Jacques (2017) develop a Gaussian Process Regression framework to predict a multidimensional poverty index at the commune level in Senegal by combining two complementary data sources: mobile phone metadata (CDRs) and diverse environmental and geospatial features [PJ17]. The CDR data capture behavioral patterns of over 9 million users based on more than 11 billion call and text records, with features such as activity levels, mobility, communication diversity, and responsiveness. The environmental data include over 40 indicators covering food security, economic activity, accessibility to services, and physical geography. Leveraging these heterogeneous and high-dimensional features, their multi-source model achieves strong predictive performance, with a spatially cross-validated Pearson correlation of 0.91 against census-derived poverty values.

Putri et al. (2023) combined two complementary data sources to estimate poverty in East Java: satellite imagery and POI data. Satellite inputs included Sentinel-2 multispectral bands, capturing land cover and infrastructure patterns, and VIIRS nighttime light intensity, which reflects economic activity and electrification [PWP23]. These were integrated with POI data from OpenStreetMap, providing fine-grained information on public facilities and services such as schools, hospitals, and markets. The study found that combining both data types significantly improved model accuracy, with the best performance achieved when multi-source features were used together. This highlights

the advantage of fusing physical and socio-economic signals for high-resolution poverty mapping.

### 2.1.2 Cross-Country Training and Transferability in Poverty Mapping

Recent studies increasingly incorporate data from multiple countries to develop combined models for poverty mapping, often employing cross-country training and application strategies. Cross-country training typically involves constructing a single model using data pooled from several countries, which is then applied to a specific target country. This approach enables models to learn from a broader distribution of poverty-related features and to capture shared socioeconomic patterns across diverse national contexts. Although single-country models generally achieve the highest performance, cross-country training can, in certain cases, enhance predictive accuracy and generalization, particularly in regions with limited data availability.

Chi et al. (2022) trained a gradient-boosted decision tree model using survey and geospatial data from 56 low- and middle-income countries to create a globally consistent wealth index [CFCB22]. They compared a global model, trained on data from all countries with leave-one-country-out cross-validation, to models trained individually for each target country. When evaluated using spatial cross-validation, the global model slightly outperformed single-country models on average ( $R^2 = 0.59$  compared to 0.56), demonstrating the benefits of cross-country training. However, when the dataset was restricted to a more homogeneous group of 23 African nations, the model explained 71% of the variation using leave-one-country-out cross-validation, underscoring the importance of country similarity, as including dissimilar countries significantly reduced performance. Additionally, models trained on one country's data performed best when applied to neighboring countries or those with similar observable characteristics.

Similarly, Lee and Braithwaite (2022) compared single- and cross-country machine learning models to estimate wealth in 25 sub-Saharan African countries, utilizing geospatial features and satellite imagery [LB22]. Cross-country performance was assessed using a leave-one-country-out strategy. On average, single-country models achieved slightly higher accuracy ( $R^2 = 88.1\%$ ) than cross-country models ( $R^2 = 85.6\%$ ). However, in four countries, the cross-country model outperformed its single-country counterpart, suggesting that cross-country training can sometimes produce more robust estimations, particularly when within-country data are limited.

Marty and Duhaut (2024) further explore the transferability of models across countries by developing a comprehensive machine learning framework to estimate wealth levels and temporal changes using data from 59 countries and over 63,000 survey clusters [MD24]. Their approach integrates diverse data sources, including nighttime luminosity, daytime satellite imagery, synthetic aperture radar, Facebook marketing insights, OpenStreetMap infrastructure, land cover data, climate, and pollution indicators. They find the highest predictive accuracy when combining all available feature types, achieving an average  $R^2$  of 55% at the cluster level and 59% at the district level. When assessing model transferability

through various training strategies, within-country models are most effective in 58% of countries. Training on geographically and socioeconomically similar countries within the same continent is optimal for 20%, whereas global leave-one-country-out training yields the best results for 19% of cases. Countries with more training data benefit most from within-country models, while global or intra-continental models perform better in data-scarce or regionally consistent settings.

Collectively, these studies emphasize the importance of country similarity in cross-country transferability. They demonstrate the value of combining data from multiple contexts to enhance model robustness and accuracy.

## 2.2 Transfer Learning for Tabular Data

Gradient-boosted decision tree (GBDT) models, such as XGBoost, CatBoost, and LightGBM are frequently selected for tasks involving tabular data due to their high performance. These models are typically trained from scratch for each specific task. In contrast, transfer learning has been highly successful in deep neural networks, where models are pretrained on large datasets to improve performance on related tasks. However, GBDT models lack inherent mechanisms to leverage multiple similar datasets effectively. To address this, specialized techniques and architectures have been developed to enable transfer learning for tabular data, offering advantages such as improved performance with limited data and enhanced robustness when handling missing data.

A straightforward approach to transfer learning is to train a model on all available data from multiple sources. However, if the data distributions differ significantly, this can lead to negative transfer, where the inclusion of dissimilar source data degrades model performance. Negative transfer occurs when training data from a source domain introduce misleading patterns not representative of the target domain. To mitigate this, strategies such as re-weighting training instances can be employed, down-weighting source-domain examples that diverge from the target distribution [SCD<sup>+</sup>22]. Alternatively, a country similarity index can be used to identify source countries with characteristics closely aligned with the target country for training.

Another technique for GBDT models is model stacking, where predictions from related models are incorporated as additional features during training. In the context of poverty mapping, this could involve using a model trained on data from Liberia to generate poverty predictions for Sierra Leone, which are then included as an input feature for the Sierra Leone model. This approach has demonstrated promising results in other domains. For instance, when applied to a medical dataset with twelve diagnostic tasks, stacking with CatBoost improved AUROC by an average of 0.05 [LCS<sup>+</sup>22].

Several GBDT architectures have been specifically adapted for transfer learning to address the challenges of applying models across domains with differing data distributions. One such approach is TransBoost, which integrates tree-based models with kernel-based domain adaptation [SLW<sup>+</sup>22]. TransBoost trains two boosting ensembles—one for the source

domain and one for the target domain—in parallel, utilizing a shared tree structure but assigning different weights to nodes. This method has demonstrated superior prediction accuracy compared to other state-of-the-art transfer learning techniques, while also being more efficient and robust against sparse features. However, a limitation of TransBoost is that it is designed exclusively for classification tasks, which limits its applicability to regression-based poverty mapping.

Another technique is TrAdaBoost, an instance-based transfer learning algorithm that extends the AdaBoost framework to handle distribution mismatches between source and target domains [DYXY07]. TrAdaBoost iteratively reweights training samples to prioritize relevant knowledge from the source domain while mitigating negative transfer. Specifically, target instances are upweighted when misclassified to emphasize their importance, whereas source instances are downweighted when misclassified, under the assumption that they may introduce misleading patterns. This approach is particularly effective in low-data regimes, making it well-suited for poverty mapping, where target country data is often scarce but related source data from other countries is available. By focusing on relevant source knowledge, TrAdaBoost can enhance model performance in data-constrained settings.

TrAdaBoost exemplifies an instance-based approach, whereas feature-based approaches aim to directly align or restructure feature spaces across domains. Correlation Alignment (CORAL) reduces distributional differences by aligning the second-order statistics of source and target feature representations through linear transformations [SFS16]. Feature Augmentation (FA) [DI07] instead expands the feature space, creating separate domain-specific and domain-general copies of each feature. This allows the model to learn which dimensions transfer across domains and which should remain context-specific.

Another research direction involves Deep Tabular Models (DTMs), which offer a more natural framework for parameter-based transfer learning than GBDTs. While GBDTs generally outperform DTMs on tabular data benchmarks [ZSE<sup>+</sup>23], certain studies demonstrate that DTMs can surpass GBDTs when pretrained on related datasets. For instance, pretraining DTMs on relevant data has been shown to yield superior performance, providing a rationale for their application in poverty map inference [LCS<sup>+</sup>22].

Popular DTMs include TabNet [AP21], TabTransformer [HKCK20], FT-Transformer [GRKB21], and XTab [ZSE<sup>+</sup>23]. Notably, the FT-Transformer has demonstrated the ability to match or outperform GBDTs, particularly in scenarios with a large number of classes or high input dimensionality [GRKB21].

A recent advancement in this field is the development of foundation models for tabular data, which involve training large-scale models on extensive collections of tables to enable generalization to new tasks with minimal training data. TABULA-8B, a transformer-based model pretrained on hundreds of millions of tabular rows using masked language modeling objectives adapted for structured data, exemplifies this approach [GPS24]. Each table row is serialized as a sequence of “feature: value” tokens, enabling robust knowledge transfer. TABULA-8B has achieved remarkable few-shot performance, outperforming

GBDT models by 5–15 percentage points in accuracy with only 1–32 labeled examples. This performance advantage persists even when baseline models are trained on datasets up to 16 times larger, underscoring the power of foundation models for tabular data in low-data settings.

Among DTMs, TransTab stands out as a model specifically designed for transfer learning on tabular datasets with varying feature sets [WS22]. TransTab introduces flexibility by embedding each feature alongside a column identifier, allowing for differences in the columns present during pretraining and fine-tuning. This removes the requirement for identical dataset structures, making TransTab especially valuable for poverty mapping, where feature availability often varies across countries. In benchmark tasks, such as clinical trial mortality prediction, TransTab has outperformed other DTMs, highlighting its potential for transfer learning in scenarios with heterogeneous data.

Handling missing features is critical for effective transfer learning, as missing data in either the source or target dataset can reduce model performance if important features are excluded. Standard approaches to handling missing data in tabular machine learning commonly include techniques such as mean or mode imputation, k-nearest neighbors (KNN) imputation, and model-based methods. These techniques are typically applied within a single dataset and do not incorporate information from external sources. In transfer learning settings, however, features that are entirely unobserved in the target domain may be reconstructed using data from the source domain. This can be done by training a predictive model on the source domain, where the feature is available, using other covariates, and then applying this model to estimate the feature in the target domain. Levin et al. (2023) use this method for missing data in upstream or downstream datasets and report consistently improved model performance compared to setups without imputation [LCS<sup>+</sup>22].

When performing cross-country imputation, it is important to assess whether the relationship between features holds across domains. If a feature has a very different distribution in the target country, a transferred imputation model might introduce bias.

## 2.3 Country Similarity Indices

Country similarity indices are quantitative measures that capture how closely two or more countries resemble each other across specified characteristics, facilitating applications in fields such as economics, sociology, and geography. These indices draw on diverse features, including economic, cultural, political, environmental, and infrastructural factors, as well as geographic proximity.

A common method involves cluster analysis using variables such as GDP per capita, literacy rates, life expectancy, and industrialization levels to group countries into clusters of comparable countries based on shared socioeconomic traits. Composite indicators, which aggregate heterogeneous factors into a single metric, have also gained prominence for their ability to capture multifaceted dimensions of similarity. A notable example

outside economics is Hofstede’s six-dimensional national culture framework, which remains a widely adopted standard for assessing cultural similarity between countries [Hof80].

Geography itself provides some of the simplest metrics for country similarity. The most straightforward is spatial distance, e.g., the great-circle distance between capitals or centroids, based on the assumption that geographically closer countries tend to be more similar due to historical and social interaction. To that end, geographers have also compared countries based on environmental and land use features. For instance, one could measure climate similarity using Köppen climate zone distributions or compare the percentage of various land cover types [BMV<sup>+</sup>23].

In urban studies and geography, similarity indices often focus on human settlements and infrastructure. Metrics such as urbanization levels, city size distributions, and infrastructure attributes, including transport networks and telecommunication coverage, provide insights into structural similarities. A recent work leverages high-dimensional urban data, such as POI data to quantify fine-grained similarities between countries [WCC<sup>+</sup>25].

Country similarity indices differ in how they define and calculate similarity, ranging from simple overlaps to statistical distances and composite measures. Each method has strengths and limitations, and results can vary depending on the choice of metric. While useful, these indices face critiques for oversimplification, arbitrary weighting, and neglect of context-specific nuances. Researchers increasingly recommend using multiple indices and validating them against meaningful outcomes [She01].

Country similarity indices can be valuable tools in machine learning, particularly in transfer learning, where they guide the effective transfer of models across domains by identifying source data with characteristics closely aligned to the target. They can help mitigate negative transfer and enhance model performance in data-scarce settings. Recent studies demonstrate their utility across diverse applications, from health modeling to natural language processing and urban analytics.

One notable application is presented by Nalmpatian et al. (2024) [NHAI24]. In the framework, a country similarity score guides the transfer of mortality models to a target country lacking local data. The score is computed using a Manhattan distance between standardized vectors of 13 external indicators, covering life insurance metrics, healthcare statistics, and population mortality rates. This distance is transformed into a similarity score, which is then used to weight the resampling of data from each source country when generating a synthetic target-country dataset. This ensures that the synthetic data used for fine-tuning the transfer model is both representative and diverse, yet grounded in similarity to the target context.

In their comprehensive analysis of factors affecting multilingual language model (MLLM) performance, Bagheri Nezhad et al. (2024) identify country similarity as a key driver of cross-lingual transfer effectiveness [BNAP24]. Unlike purely linguistic or geographic measures, country similarity captures shared sociocultural and political contexts between languages by quantifying the overlap of countries in which the languages are spoken,

using a Jaccard similarity index. Interestingly, the results showed that country similarity was among the top predictors of model performance, more so than simple geographic proximity.

Wang et al. (2025) propose a similarity-based cross-city transfer learning framework [WCC<sup>+</sup>25]. The core idea is to improve transfer learning performance by ensuring that knowledge is only transferred between cities with similar urban characteristics. They construct a city similarity model based on 19-dimensional POI (Point of Interest) data, capturing urban functional structures across categories such as education, shopping, and transportation. Cities are then clustered using k-means, and only those in the same cluster are used for transfer. This clustering prevents negative transfer from dissimilar cities. Ablation studies show the clustering step boosts predictive accuracy compared to unconstrained transfer.

These studies underscore the relevance of similarity indices in enhancing the effectiveness of transfer learning.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# CHAPTER 3

## Data

This chapter describes the data used for the analysis, including the selection of countries, the ground-truth wealth data, and the geospatial feature sources. It details the construction of the ground-truth wealth indicator based on household survey data. Then, it presents the external feature datasets that provide information on infrastructure, economic activity, population distribution, and connectivity, which are used as predictors in the modeling process.

### 3.1 Selection of Six Sub-Saharan African Countries for Analysis

Sierra Leone and Uganda were initially selected due to their frequent inclusion in previous literature (e.g. [ENKK23]). To enhance the diversity of the analysis and to test the hypothesis from existing literature suggesting that neighboring countries may perform well in model transfer scenarios, Liberia and Rwanda, neighboring Sierra Leone and Uganda respectively, were included. To further broaden the socioeconomic spectrum and assess model performance across varying economic contexts, Gabon and South Africa, which are relatively wealthier Sub-Saharan African nations, were also selected. This selection creates a balanced sample with varying development levels and regional clusters.

The geographic locations of these six countries are illustrated in Figure 3.1.

### 3.2 Ground-truth Data

The ground-truth data for this study originates from nationally representative household-level surveys conducted by the Demographic and Health Surveys Program [The22], specifically the Standard Demographic and Health Survey (DHS) and the Malaria Indica-

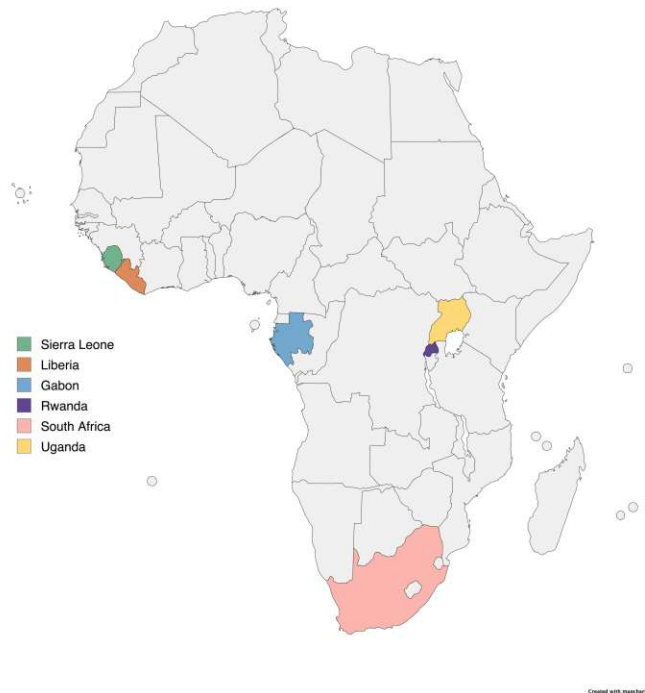


Figure 3.1: Geographic locations of the Sub-Saharan African countries analyzed: Sierra Leone, Uganda, Liberia, Gabon, South Africa, and Rwanda.

tor Survey (MIS). Both surveys capture detailed information on housing characteristics and asset ownership, enabling the estimation of household wealth.

**Cluster Locations.** Survey respondents are grouped into clusters, which typically correspond to census enumeration areas selected proportionally to their size within each stratum. Each cluster usually contains 25–30 households and is classified either as urban or rural. To preserve respondent anonymity, clusters are geographically displaced by up to 2 km in urban areas and up to 5 km in rural areas, with 1% of rural clusters displaced by up to 10 km.

**International Wealth Index (IWI).** The International Wealth Index (IWI) provides a standardized, asset-based measure of households’ material well-being that is comparable across low- and middle-income countries. It is derived from a consistent subset of asset ownership and housing characteristics questions included in DHS and MIS surveys. Principal component analysis (PCA) is applied to asset ownership and housing characteristics, generating weights that produce an index ranging from 0 to 100, with 0 indicating the lowest and 100 the highest economic status. For each cluster, household-level IWI scores are averaged, producing a mean and a standard deviation for the cluster-level IWI scores used as ground-truth for modeling.

Table 3.1 reports the number of census clusters available for each country.

Country	Number of Clusters
Sierra Leone	893
Uganda	1001
Liberia	471
Gabon	390
South Africa	746
Rwanda	500

Table 3.1: Number of clusters per country

Figure 3.2 presents the distribution of the wealth index for each selected country, highlighting variation in economic conditions. These distributions reveal underlying economic differences that are important for understanding the challenges of cross-country model adaptation.

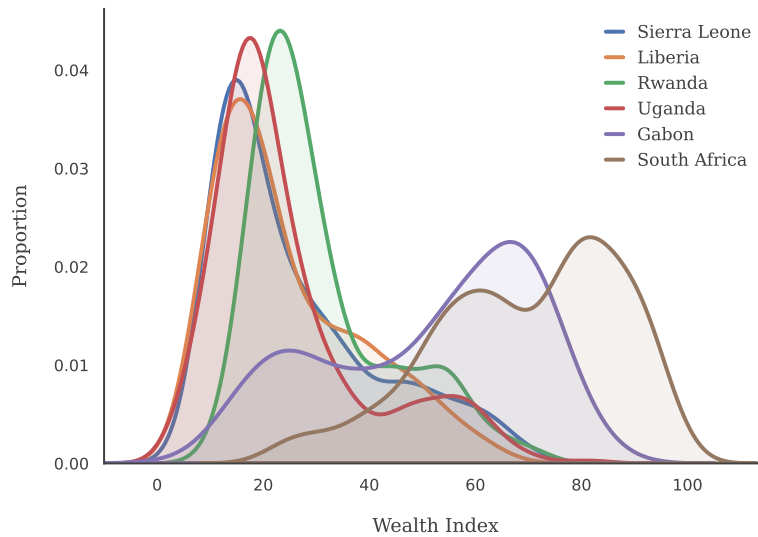


Figure 3.2: Distribution of cluster-level mean IWI for six countries. Curves are Gaussian kernel-density estimates, so the area under each curve equals 1.

### 3.3 Feature Sources

In our modeling approach, we utilized a diverse set of features derived from multiple geospatial data sources to capture aspects of infrastructure, population distribution, and connectivity.

#### OpenStreetMap (OSM) Features

OpenStreetMap is a collaborative project that provides freely accessible geospatial data collected by a community of contributors. From OSM, we extracted two categories of

features:

**Type of Settlement (1 variable):** This feature differentiates between urban and rural areas. It is constructed from OSM classifications, which categorize places as cities, towns, neighborhoods (urban), villages, hamlets, and isolated dwellings (rural).

**Infrastructure Features (54 variables):** Counts and distances of roads, buildings, and 24 categories of points of interest, such as schools, health centers, banks and markets, within radii of 1.6–10 km. These features reflect both availability and accessibility of essential services.

#### **Nighttime Light-Based Features (NTL) (36 variables)**

Nighttime light data using VIIRS-DNB imagery, obtained from the Google Earth Engine, serves as a proxy for economic activity and development. These data provide globally consistent measurements of artificial illumination at night, which strongly correlates with electrification, infrastructure density and urbanization. Nightlight intensity is a well-established feature in poverty mapping studies [YPD<sup>+</sup>20, JBX<sup>+</sup>16].

A total of 36 features were derived by summarizing light intensity within buffer zones of 1.6, 2.0, 5.0, and 10.0 km. These features include standard statistics (minimum, maximum, mean, median), fractional metrics such as the proportion of lit pixels, cumulative metrics like total radiance, and temporal aggregates including the trailing 30-day average and long-term mean intensity. High feature values typically correspond to urban centers, while rural regions often exhibit consistently low or near-zero radiance levels.

#### **Facebook Population (FBP) Features (9 variables)**

High-resolution population density maps produced by Meta's "Data for Good" initiative were used to construct nine features related to population distribution. These maps estimate the number of people living within 30-meter grid tiles globally, providing detailed insights into settlement patterns and population density.

The features include the distance to the nearest populated tile, the population value in the closest tile and the total population within buffer zones of 1.6, 2.0, 5.0, and 10.0 km. Gravitational accessibility metrics were also calculated, combining population and distance following the formula  $\text{population}/\text{distance}^\beta$  with  $\beta \in \{1.0, 1.5, 2.0\}$ . This approach captures the idea that larger nearby populations contribute more strongly to local accessibility than distant ones.

#### **OpenCellID Connectivity Features (OCI) (9 variables)**

OpenCellID is a global crowdsourced database that aggregates GPS locations and attributes of mobile cell towers. From this dataset, features were constructed to capture aspects of mobile network connectivity. These include the distance to the nearest cell tower, the number of cell towers and unique mobile cells within buffer zones of 1.6, 2.0, 5.0, and 10.0 km.. These features reflect the availability and density of mobile network infrastructure.

# Methods

This chapter presents the methodological foundations of the study. It describes the machine learning architectures used and the experimental setup for model evaluation. Additionally, the chapter introduces a suite of country similarity indices designed to guide model transfer by quantifying socioeconomic, geographic, and cultural alignments between countries.

## 4.1 Machine Learning Architectures

This section details the machine learning models selected for the prediction task. Three distinct architectures are evaluated: CatBoost, a powerful gradient boosting framework for tabular data; TrAdaBoostR2, an algorithm for instance-based transfer learning; and TransTab, a transformer-based approach designed for flexible tabular data modeling. These choices represent a range of techniques suitable for the data characteristics and transfer learning objectives of this study.

### 4.1.1 CatBoost

CatBoost is a gradient boosting framework optimized for handling categorical variables effectively, thus making it particularly suitable for tabular datasets with diverse feature types [PGV<sup>+</sup>18]. The algorithm iteratively builds an ensemble of decision trees that minimize a specified loss function by employing gradient descent in function space. Unlike other gradient boosting frameworks, CatBoost incorporates several techniques specifically designed to improve generalization and reduce overfitting, particularly with categorical features.

**Ordered Target-Based Encoding.** To encode categorical features, CatBoost replaces each categorical value with a target statistic, typically the mean of the target variable for all training samples sharing the same category. This approach captures useful information

about the relationship between a category and the outcome, especially for high-cardinality features. However, calculating target statistics over the entire training set introduces prediction shift: the model indirectly uses a sample's true target when encoding its own features, leading to a subtle bias and a risk of overfitting. Specifically, if a rare category appears only a few times, its target statistic could exactly match the sample's target, allowing the model to memorize training data rather than generalize. CatBoost solves this problem by computing ordered target statistics: for each sample, it uses only information from samples that precede it in a random permutation of the training data. This ensures that the target value of a sample does not influence its own encoding, thus eliminating the leakage that causes prediction shift.

CatBoost employs additional regularization strategies. Shrinkage of the learning rate reduces the impact of each tree added to the model, slowing down learning to improve generalization. Early stopping halts training when performance on a validation set no longer improves, avoiding overfitting to the training data. CatBoost constructs ensembles of oblivious decision trees, which differ from classical decision trees in that each tree uses the same splitting criterion across all nodes at a given depth. This symmetry significantly speeds up computations and reduces model complexity, improving generalization.

**Gradient Boosting Procedure.** At each iteration  $t$ , CatBoost builds a new tree to fit the negative gradient of the loss function:

$$f_t(x) = f_{t-1}(x) + \eta \sum_{j=1}^J \gamma_j \mathbb{I}(x \in R_j) \quad (4.1)$$

where  $f_{t-1}(x)$  is the model constructed up to iteration  $t - 1$ ,  $\eta$  is the learning rate,  $J$  is the number of leaves,  $R_j$  denotes the regions defined by the leaves, and  $\gamma_j$  is the optimal leaf weight minimizing the loss function. CatBoost employs a second-order approximation to efficiently compute leaf weights and splits, further improving accuracy and stability.

**MultiRMSE Loss Function.** The prediction of two target variables, the cluster mean and standard deviation of the wealth index, necessitates a loss function suitable for multi-output regression. The MultiRMSE loss function, implemented in CatBoost, is employed for this purpose. MultiRMSE extends the standard Root Mean Squared Error (RMSE) to handle multiple target variables by computing the RMSE for each of the  $K$  target dimensions independently and then averaging these values to produce a single loss metric.

The formula is:

$$\text{MultiRMSE} = \frac{1}{K} \sum_{k=1}^K \sqrt{\frac{1}{N} \sum_{i=1}^N (y_{ik} - \hat{y}_{ik})^2}$$

where  $K$  is the number of targets,  $N$  is the number of samples,  $k$  indexes targets,  $i$  indexes samples,  $y_{ik}$  is the actual value and  $\hat{y}_{ik}$  is the predicted value.

Due to these specialized strategies, CatBoost achieves superior performance on datasets with categorical features and has demonstrated robustness in various tabular data prediction tasks, making it suitable for this study’s task of poverty inference using multimodal data.

**Feature Importance.** Feature importance is quantified directly from the structure of the trained CatBoost ensemble by examining how each split on a given feature alters the model’s output. CatBoost supports two principal metrics, `PredictionValuesChange` and `LossFunctionChange`. `PredictionValuesChange` works by recording every split on a feature, measuring the absolute change in the raw prediction at that split, weighting it by the number of samples passing through, and then summing and normalizing across all trees to yield a relative importance score. In contrast, `LossFunctionChange` estimates how much the overall loss would increase if a feature’s splits were removed: it simulates removal by discarding those splits, recomputing affected leaf values via weighted averages, and computing the loss difference on held-out data. Because the objective is MultiRMSE (non-ranking), feature importance is computed with CatBoost’s default `PredictionValuesChange`.

#### 4.1.2 TrAdaBoostR2

TrAdaBoostR2 (Transfer AdaBoost) is an instance-based transfer learning algorithm designed to address scenarios where the marginal distributions of source and target domains differ, yet both share the same feature and label space [DYXY07]. This makes it particularly suitable for problems such as poverty mapping across countries, where target domain data (e.g., from a low-surveyed region) is limited but similar source data from related regions is available.

AdaBoost is an iterative algorithm that builds a strong classifier as a weighted combination of weak classifiers. In each iteration, AdaBoost trains a weak classifier on a weighted version of the training data, assigning higher weights to instances misclassified in previous iterations. TrAdaBoostR2 adapts AdaBoost for transfer learning with one source domain and one target domain. The key idea behind TrAdaBoostR2 is to iteratively adapt the contribution of source domain samples during training. The method is based on a “reverse boosting” principle. TrAdaBoostR2 maintains separate weights for source and target instances. Like AdaBoost, it increases the weights of misclassified target instances. However, it decreases the weights of misclassified source instances, based on the assumption that source examples consistently misclassified by learners focused on the target are likely drawn from parts of the source distribution irrelevant or detrimental to the target task.

The MultiSource-TrAdaBoostR2 algorithm is used [YD10], which is an extension of the TrAdaBoostR2 framework specifically designed to leverage multiple source domains for transfer learning. The core strategy of MultiSource-TrAdaBoostR2 involves iteratively selecting the most beneficial source domain from a pool of multiple candidates based on their current relevance to the target domain. The algorithm begins by initializing a

combined weight vector for all available instances, assigning initial weights separately to each source domain and the target domain. During each iteration, the algorithm independently computes candidate weak classifiers from each source domain combined with the target domain. Each candidate classifier’s performance is evaluated based on its classification error specifically on the target domain data. Subsequently, the candidate classifier with the lowest error is selected to update the ensemble classifier. By continuously selecting the most suitable source at each iteration, it effectively identifies and exploits beneficial knowledge across multiple domains. Empirical evaluations indicate increased performance compared to TrAdaBoostR2, along with improved robustness reflected by less variation across multiple source domains [YD10].

### 4.1.3 TransTab

Traditional machine learning models for tabular data often require a fixed data schema with identical column sets for both training and inference. TransTab (Transferable Tabular Transformer) is a transformer-based deep learning model designed to handle varying table structures [WS22]. It enables learning and generalization across multiple tables within a domain, even when the available columns differ. This section outlines the core components of the TransTab architecture.

**Input Processing.** The input processor converts tabular data into semantically encoded token sequences. Instead of relying solely on cell values and fixed column indices, TransTab contextualizes cell information using column metadata, typically the column name or description. The processor handles different feature types as follows:

- **Categorical and textual features:** The column name and the cell’s textual value are concatenated. This combined string is then tokenized and each token is mapped to its corresponding embedding vector.
- **Binary features:** These are included only if their value is true, in which case the column name is tokenized and embedded.
- **Numerical features:** The column name is tokenized and embedded to capture its semantic meaning. The numerical value then scales the column embedding (scalar–vector multiplication). Scaling the semantic embedding of the column by the feature’s value aims to better preserve the feature’s magnitude information.

This input processing method allows TransTab to interpret features based on their semantic meaning rather than their fixed position, facilitating knowledge transfer even when column names or value representations differ slightly across tables.

**Gated Transformer Layers.** TransTab uses stacked gated transformer layers. Each layer refines the representations by modeling interactions between the different feature tokens. A standard Transformer layer typically consists of a multi-head self-attention mechanism followed by a position-wise feedforward network. TransTab replaces the

standard feedforward network with a gated variant. Following the attention mechanism, a token-wise gating layer controls the magnitude of each token embedding before further processing. This gating mechanism allows the model to dynamically emphasize or suppress information from specific tokens based on the learned context.

**Learning Modules: Supervised and Contrastive Pretraining.** TransTab supports both supervised and self-supervised training strategies.

- **Supervised Learning:** The classification [CLS] token is used by a classifier or regressor to make predictions for the target variable.
- **Vertical-Partition Contrastive Learning (VPCL):** The authors propose VPCL enabling pretraining across heterogeneous tables.
  - *Self-supervised VPCL:* Self-supervised VPCL operates on unlabeled data. For a given sample, multiple vertical partitions are created by taking different subsets of its columns. Partitions derived from the same sample are treated as positive pairs, while partitions from different samples serve as negative pairs. The model is trained using a contrastive loss to pull representations of positive pairs closer and push negative pairs apart in the embedding space. This encourages the model to learn representations invariant to the specific subset of columns present.
  - *Supervised VPCL:* When labels are available during pretraining, Supervised VPCL extends the contrastive objective. Partitions derived from any samples belonging to the same class are considered positive pairs, while partitions from samples of different classes form negative pairs. This leverages label information to learn more discriminative representations compared to Self-VPCL or standard supervised pretraining, potentially offering better transferability and robustness.

#### 4.1.4 Correlation Alignment (CORAL)

Correlation Alignment (CORAL) [SFS16] is a domain adaptation method designed to minimize domain discrepancies by aligning the second-order statistics (covariances) of the source and target feature distributions. This technique has demonstrated effectiveness in reducing domain shifts, thereby improving model generalization when applying trained models to data from different domains, as in cross-country poverty mapping scenarios.

Formally, let the source domain dataset be represented by  $D_S = \{x_i^S\}_{i=1}^{n_S}$  and the target domain dataset by  $D_T = \{x_j^T\}_{j=1}^{n_T}$ , where  $x_i^S, x_j^T \in \mathbb{R}^d$  are  $d$ -dimensional feature vectors.

CORAL aims to find a linear transformation  $A \in \mathbb{R}^{d \times d}$  that aligns the covariance matrix  $C_S$  of the source domain with the covariance matrix  $C_T$  of the target domain. The covariance matrices for the respective domains are computed as:

$$C_S = \frac{1}{n_S - 1} (D_S^T D_S - \frac{1}{n_S} (\mathbf{1}^T D_S)^T (\mathbf{1}^T D_S)) \quad (4.2)$$

$$C_T = \frac{1}{n_T - 1} (D_T^T D_T - \frac{1}{n_T} (\mathbf{1}^T D_T)^T (\mathbf{1}^T D_T)), \quad (4.3)$$

where  $\mathbf{1}$  denotes a column vector of ones.

To align the domains, CORAL optimizes the following objective function:

$$\min_A \|A^T C_S A - C_T\|_F^2, \quad (4.4)$$

where  $\|\cdot\|_F$  is the Frobenius norm.

A closed-form solution for this transformation  $A$  is achieved by whitening the source domain data followed by re-coloring using the covariance structure of the target domain:

$$A = C_S^{-\frac{1}{2}} C_T^{\frac{1}{2}}. \quad (4.5)$$

In practice, a regularization term  $\lambda I$  is added to ensure numerical stability when inverting covariance matrices:

$$A = (C_S + \lambda I)^{-\frac{1}{2}} (C_T + \lambda I)^{\frac{1}{2}}, \quad (4.6)$$

where  $I$  is the identity matrix and  $\lambda$  is a small regularization parameter.

After transformation, the source features become aligned with the target domain, significantly reducing distributional discrepancies. CORAL is particularly beneficial in poverty mapping tasks across countries, where datasets often differ due to diverse socioeconomic conditions, data collection methodologies, and geographic disparities. By aligning the covariance structures, CORAL enhances the performance and transferability of predictive models trained on heterogeneous sources.

#### 4.1.5 Feature Augmentation (FA)

Feature Augmentation (FA) [DI07] is a domain adaptation technique that addresses domain differences by augmenting the original feature space to explicitly capture domain-specific and domain-general information. This simple yet effective approach enables standard supervised learning algorithms to inherently manage domain discrepancies, thus facilitating transfer learning across diverse domains, such as poverty mapping tasks involving multiple countries.

Formally, let the original feature space be denoted as  $X \subseteq \mathbb{R}^F$ , where  $F$  is the number of original features. Given a source dataset  $D_S$  and a target dataset  $D_T$ , FA constructs an augmented feature space  $\tilde{X}$  defined as follows:

$$\tilde{X} = \mathbb{R}^{3F}, \quad (4.7)$$

$$\Phi_S(x) = \langle x, x, \mathbf{0} \rangle, \quad (4.8)$$

$$\Phi_T(x) = \langle x, \mathbf{0}, x \rangle, \quad (4.9)$$

where  $\mathbf{0} \in \mathbb{R}^F$  denotes a zero vector,  $\Phi_S$  is the mapping for source data points, and  $\Phi_T$  is the mapping for target data points.

The augmented feature representation in Feature Augmentation (FA) explicitly distinguishes between general, source-specific, and target-specific information. Rather than assuming that all features have the same meaning across domains, FA creates parallel versions of each feature, allowing the model to learn where a feature is broadly transferable and where its relevance is confined to a particular domain. For example, a feature correlated with poverty in one country might carry a different implication in another. FA makes this distinction explicit, enabling standard supervised algorithms to identify which aspects generalize across domains and which should remain domain-specific.

In the multi-domain scenario (e.g., poverty mapping across multiple countries), FA generalizes naturally by creating  $K + 1$  copies of each feature, where  $K$  is the number of domains. Specifically, the augmented feature space becomes:

$$\tilde{X} = \mathbb{R}^{(K+1)F}, \quad (4.10)$$

where each domain has its dedicated feature subset plus a common general feature subset.

FA can also be kernelized, leading to a modified kernel function  $\tilde{K}(x, x')$  defined as:

$$\tilde{K}(x, x') = \begin{cases} 2K(x, x'), & \text{if } x, x' \text{ from same domain} \\ K(x, x'), & \text{if } x, x' \text{ from different domains} \end{cases} \quad (4.11)$$

where  $K(x, x')$  denotes the original kernel function. This kernel interpretation suggests that data points from the same domain are treated as inherently more similar, reinforcing the domain structure in the feature space.

Due to its simplicity and effectiveness, FA is particularly suited for complex tasks such as cross-country poverty mapping, where domain differences are prevalent, and model generalization across domains is essential.

## 4.2 Experimental Setup

To ensure robust and reliable evaluation of model performance, all experiments were conducted using five-fold cross-validation. The reported performance metrics represent

the average results across the five folds. To enable comparability across methods and countries, the same cross-validation splits were consistently applied within each country.

Hyperparameter optimization is conducted for each method on the training set to ensure a fair comparison across methods. Bayesian optimization, a probabilistic approach that efficiently searches the hyperparameter space by modeling the objective function (e.g., validation loss) as a Gaussian process, is employed. Unlike grid or random search, Bayesian optimization balances exploration and exploitation by leveraging prior evaluations to predict promising hyperparameter configurations, thereby reducing computational cost and improving efficiency. For each method, optimization is performed using four-fold cross-validation on the training set to ensure robust parameter selection. Once optimal hyperparameters are identified, a final model is trained on the entire training set for each split and performance is evaluated on the independent test sets.

For the CatBoost model, the optimized hyperparameters include the number of iterations, tree depth, learning rate, and L2 leaf regularization strength, which control model complexity and prevent overfitting. For TrAdaBoostR2, the tuned parameters are the number of iterations, tree depth, minimum number of data points required at a node before splitting, and the minimum number of data points required in each leaf after a split, ensuring stable and effective adaptation to the target domain. For TransTab, a deep tabular model designed for transfer learning, the optimized hyperparameters include the number of training epochs, learning rate, and batch size, which most strongly influence the training dynamics and convergence of the neural network.

#### 4.2.1 Evaluation Metrics

Model performance is primarily evaluated using a normalized version of the Root Mean Squared Error (NRMSE). To ensure comparability with previous studies and to provide an easily interpretable measure, the coefficient of determination ( $R^2$ ) is additionally reported, quantifying the proportion of variance explained by the model.

**Normalized Root Mean Squared Error (NRMSE).** The NRMSE provides a measure of the average prediction error relative to the scale of the target variable. It is calculated by first computing the Root Mean Squared Error (RMSE), which represents the square root of the average squared prediction errors, and then normalizing this value by the standard deviation of the true target values  $\sigma_y$ . Lower NRMSE values indicate better model performance. It is defined as:

$$\text{NRMSE} = \frac{\text{RMSE}}{\sigma_y} = \frac{\sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}}{\sqrt{\frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{y})^2}}$$

where  $y_i$  is the true value,  $\hat{y}_i$  is the predicted value,  $\bar{y}$  is the mean of the true values, and  $N$  is the number of samples. An NRMSE of 1 indicates that the average prediction error is equal to the standard deviation of the target variable.

**Coefficient of Determination ( $R^2$ ).** The  $R^2$  metric measures the proportion of variance in the target variable explained by the model. It is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2}$$

## 4.3 Country Similarity Indices

The transferability of machine learning models across different geographical regions is a critical consideration when developing predictive models for socioeconomic outcomes such as poverty. A fundamental hypothesis is that models trained on data from one country may perform better when transferred to similar countries. To test this hypothesis, robust quantitative measures of similarity between countries are required. This section introduces a comprehensive suite of distance metrics designed to capture various dimensions of country similarity, which will be used to evaluate their effectiveness in predicting model transfer performance. These distance metrics serve as proxy measures for the underlying factors that may influence how well a machine learning model trained on one country's data generalizes to another country.

### 4.3.1 Statistical Measures

#### Wasserstein Distance

Several of the distance measures used are based on the Wasserstein distance, also known as the Earth Mover's Distance. This metric quantifies the cost of transforming one probability distribution into another, and is particularly well-suited for comparing distributions in a meaningful and interpretable way. In the one-dimensional case, which is used throughout this work, the first Wasserstein distance between two probability distributions  $P$  and  $Q$  with respective cumulative distribution functions  $F_P$  and  $F_Q$  is defined as:

$$W_1(P, Q) = \int_{-\infty}^{\infty} |F_P(x) - F_Q(x)| dx$$

This can be intuitively understood as the minimum amount of "effort" required to transform one distribution into the other when both are represented as piles of probability mass along the real line. In practice, this quantity is approximated using empirical distributions derived from the data.

#### Jaccard Index

The Jaccard Index is a measure of similarity between two sets, defined as the size of the intersection divided by the size of the union. It is particularly useful when comparing binary or categorical attributes, such as top-k feature sets or shared geographic and historical traits.

Given two sets  $A$  and  $B$ , the Jaccard Index  $J(A, B)$  is defined as:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

The corresponding Jaccard distance, which measures dissimilarity, is simply  $1 - J(A, B)$ .

### Kolmogorov–Smirnov (KS) Test

The Kolmogorov–Smirnov (KS) test is a non-parametric statistical test that compares two empirical cumulative distribution functions (CDFs) to determine whether they are drawn from the same distribution. It is particularly well-suited for comparing one-dimensional distributions.

Given two samples with empirical CDFs  $F_n(x)$  and  $G_m(x)$ , the KS statistic  $D_{n,m}$  is defined as:

$$D_{n,m} = \sup_x |F_n(x) - G_m(x)|$$

Here,  $\sup$  denotes the supremum, or maximum difference, across all values of  $x$ . The KS statistic captures the largest vertical distance between the two CDFs.

### 4.3.2 Description of Similarity Metrics

1. **PCA-Based Wasserstein Distance of Features.** The first similarity index involves dimensionality reduction via Principal Component Analysis to transform the feature data into principal components. Subsequently, the Wasserstein distance is computed between each pair of countries, capturing underlying structural differences in the features.

Two variants are defined including:

- a) The first principal component
- b) The first two principal components

2. **Feature Importance Ranking Similarity.** This index measures similarity based on feature importance rankings derived from the models in each country. Two variants exist:

- a) Kendall’s Tau: Calculating correlation across complete feature rankings to quantify alignment or divergence.
- b) Jaccard Distance: Comparing the top-k ranked features between countries to assess overlap.

3. **Facebook Movement Distribution.** Using Facebook mobility data, this distance metric applies Wasserstein distance on the distribution of human movements within and between geographic regions. This method assumes that similarities in large-scale mobility patterns might reflect underlying similarities in economic structure, infrastructure, or social connectivity relevant to poverty distribution.
4. **Polycentricity of Settlements.** This distance metric quantifies the similarity between countries based on their internal urban structure, specifically focusing on the degree of mono-centricity versus poly-centricity. Polycentricity is commonly linked to balanced economic development, diversified activities, and reduced spatial inequality, while mono-centricity indicates concentrated economic resources within a primary city. The calculation relies on analyzing the relationship between the population size of administrative units and their rank within the country. The analysis combines two publicly available geographic datasets. First, high-resolution population estimates are drawn from the Facebook High Resolution Settlement Layer. Second, the spatial definitions of administrative regions are taken from the geoBoundaries database. For each country, the total population within each ADM2 region is computed and the regions are then ranked in descending order of population size. From this relationship, several quantitative features are extracted to profile each country's population structure.
  - The slope and intercept derived from a linear fit to the log-log rank-size data, which indicates the steepness of the urban hierarchy. The largest region is excluded.
  - The deviation of the largest region's population from this fitted trend is calculated, serving as an indicator of primacy.
  - Basic statistics, such as the percentage of the total population living in the largest, smallest, and median-ranked regions.
  - Power law fit features that describe how well the observed population distribution conforms to a mathematical power law model. Here, power-law distributions are fitted to the region-level population data. The resulting parameters are the power law exponent alpha and its uncertainty. Additionally, parameters from the truncated power law fit are alpha and the cutoff parameter.

Four variants are defined including different sets of the extracted features:

- a) Includes slope, intercept, largest region deviation, and population statistics of largest, smallest, median.
- b) Includes only the power law fit features (alpha).
- c) Includes the largest region deviation, population statistics and the power law fit features.
- d) Includes all features.

5. **Rural Population Density Distribution.** This metric compares countries based on the statistical distribution of population density characteristics specifically within their rural areas. It utilizes the two-sample Kolmogorov-Smirnov (KS) test to quantify the difference between these rural density distributions for pairs of countries. The underlying assumption is that similarities in rural settlement patterns reflect shared characteristics in agricultural structure, infrastructure access, or land use relevant to rural development and poverty. Four variants are calculated, differing in the specific density feature used and whether the feature's values are standardized before comparison:

- a) Compares distributions using raw values of a highly localized population density measure.
- b) Compares distributions using standardized values within each country of the highly localized population density measure. Similarity depends primarily on the relative shape of the distribution, irrespective of absolute density levels.
- c) Compares distributions using raw values of population density aggregated within a small neighborhood radius.
- d) Compares distributions using standardized values of population density aggregated within a small neighborhood radius.

6. **Historical and Geographic Context Similarity.** This metric assesses the similarity between countries based on shared historical and geographic characteristics. It assumes that countries with similar colonial pasts and regional placement might share contextual factors relevant to development or poverty patterns. The calculation aggregates similarity across three dimensions:

- **Shared Colonizers:** Measures the overlap in the sets of historical colonizing powers.
- **Shared Subregion:** Assigns similarity if two countries belong to the same predefined geographic subregion.
- **Shared Neighbors:** Measures the overlap in the sets of bordering countries.

The Jaccard Index is computed to measure the overlap for each attribute. These individual similarity scores are then summed to produce a final metric.

7. **Jones Country Similarity Index.** The Jones Country Similarity Index [Jon25] developed by Jeff M. Jones evaluates country similarity by combining five equally weighted domains: demographics, culture, politics (government and public policy), infrastructure/technology, and geography. Within each domain, multiple subcomponents (e.g., language, religion, age structure, writing systems; regime and policy characteristics; transport, water, energy, communications, health infrastructure; climate, land cover, location, topography, hydrology) are compiled to produce domain scores, which are then averaged to obtain an overall similarity score for

each country pair. The methodology is applied uniformly across countries. This metric provides an integrated, holistic measure of similarity between nations.

8. **Conflict Profile Similarity.** This metric quantifies the similarity between countries based on their recent internal conflict patterns, utilizing data from the Armed Conflict Location & Event Data Project (ACLED) [RLHK10]. The core idea is that similarities in conflict events may reflect underlying similarities in political stability or social tensions relevant to development and poverty. For each country, conflict events are aggregated by type (e.g., Battles, Riots, Violence against civilians). These counts are then normalized to create a probability distribution representing the proportional profile of conflict types within that country. The Wasserstein distance is calculated between these conflict profile distributions for each pair of countries.
9. **GDP-Based Economic Similarity.** This metric includes two variants:
  - a) **Single-Year GDP Comparison:** Calculates relative differences in GDP per capita during the model training year.
  - b) **Multi-Year GDP Profile:** Uses Wasserstein distance to measure differences across GDP trends over multiple years, capturing economic stability and growth trajectories.
10. **National Cultural Dimensions.** This index measures similarity based on national cultural characteristics, drawing upon Hofstede's established multi-dimensional framework [Hof80]. Countries are profiled using their scores on dimensions such as power distance, individualism versus collectivism, masculinity versus femininity, uncertainty avoidance or long-term orientation. Each country is represented by a vector containing its scores across these dimensions. The similarity between pairs of countries is then assessed by calculating the first Wasserstein distance between their respective cultural profile vectors



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# Results

This chapter presents the findings from the application of machine learning models for poverty mapping across six Sub-Saharan African countries. It begins by evaluating the predictive performance of within-country CatBoost models and analyzing feature importance. The chapter then explores cross-country model transfer, assessing generalizability and the role of country similarity indices in predicting transfer success. Multi-country training and strategies for handling missing features are examined to address data scarcity challenges. Finally, transfer learning techniques are compared against within-country baselines to evaluate their effectiveness.

## 5.1 Poverty Map Models

CatBoost models are first trained separately on each country. This provides both a set of models for cross-country application and a within-country performance benchmark against which model transfer can be evaluated.

### 5.1.1 Predictive Performance

Table 5.1 summarizes the predictive performance of each model for the cluster-level wealth index mean and standard deviation, reporting the Normalized Root Mean Squared Error (NRMSE) and the coefficient of determination ( $R^2$ ).

Overall, the models predict the cluster means with relatively high accuracy, while performance for the standard deviations is notably lower. This pattern holds across all countries. South Africa represents a particular outlier, where predictions of the cluster mean are substantially less accurate compared to other countries. Since the cluster mean provides the most stable signal, the remainder of the chapter focuses on reporting results for the mean.

Table 5.1: Model results for wealth index cluster mean and standard deviation predictions.

Country	Cluster Mean		Cluster Std Dev	
	NRMSE	R <sup>2</sup>	NRMSE	R <sup>2</sup>
Sierra Leone	0.43	0.82	0.77	0.40
Liberia	0.52	0.77	0.85	0.27
Rwanda	0.58	0.68	0.90	0.19
Uganda	0.47	0.78	0.82	0.32
Gabon	0.41	0.83	0.87	0.24
South Africa	0.74	0.46	0.87	0.24

### 5.1.2 Feature Importance

**Feature importance distribution for CatBoost models.** Figure 5.1 illustrates the feature importance distributions for Uganda and Sierra Leone, ranked from most to least significant. Both countries show a steep decline in feature importance, indicating that a small subset of features disproportionately influences predictive performance.

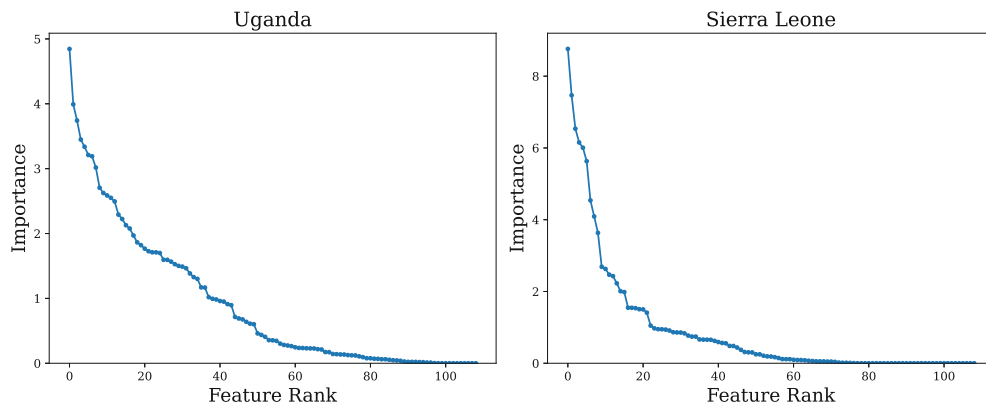


Figure 5.1: Feature importance distribution for Uganda and Sierra Leone.

**Feature category importance.** Figure 5.2 summarizes the overall importance of feature categories used in the modeling, showing both summed and average importance scores. Nighttime features contribute the highest total importance. In contrast, Facebook Population features exhibit the highest average importance per feature, suggesting these few population-based features are individually highly influential for poverty estimation.

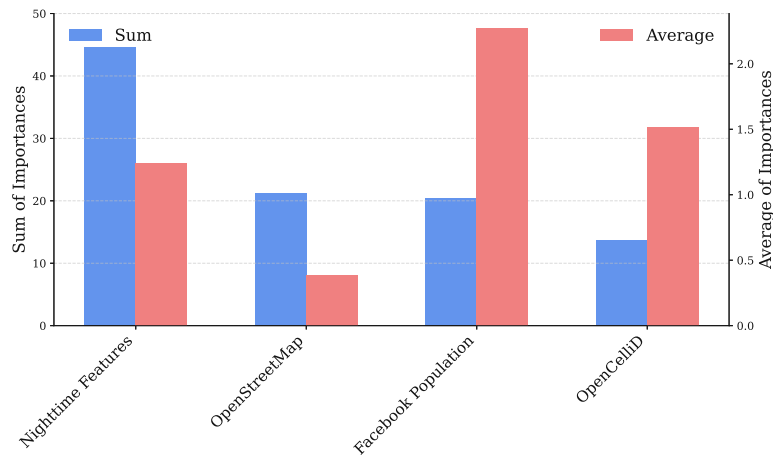


Figure 5.2: Overall sum and average importance of feature categories over all countries.

## 5.2 Model Transfer

To assess cross-country generalization, this section evaluates the transferability of models by applying those trained in one country (source) to all other countries (target). The performance of these transferred models was compared against a within-country benchmark model, trained and evaluated solely on the target country’s data.

Figure 5.3 presents the results for each target country. Across all cases, models trained on the target country outperform transferred models. However, the magnitude of performance degradation varies considerably between country pairs. Among Sierra Leone, Liberia, Rwanda, and Uganda, cross-country transfer results in only minor reductions in predictive accuracy, suggesting a relatively high degree of generalizability. In contrast, Gabon and South Africa stand out as exceptions, where none of the source models achieve performance exceeding the baseline of always predicting the mean ( $\text{NRMSE} > 1$ ). This aligns with earlier observations of substantial differences in the wealth index distributions in these two countries (cf. Figure 3.2).

From the cross-country transfer results, we identify the best-performing source country for each target. Table 5.2 summarizes these findings. Notably, Sierra Leone emerges as the most frequently effective source country, particularly for transfers within the group of West and East African countries.

### 5.2.1 Evaluating Similarity Metrics for Optimal Model Transfer Performance

The results reveal substantial variability in the success of cross-country model transfer. While models trained in certain source countries generalize well to specific target countries, performance degradation can be dramatic in other cases. Given this variability in transfer

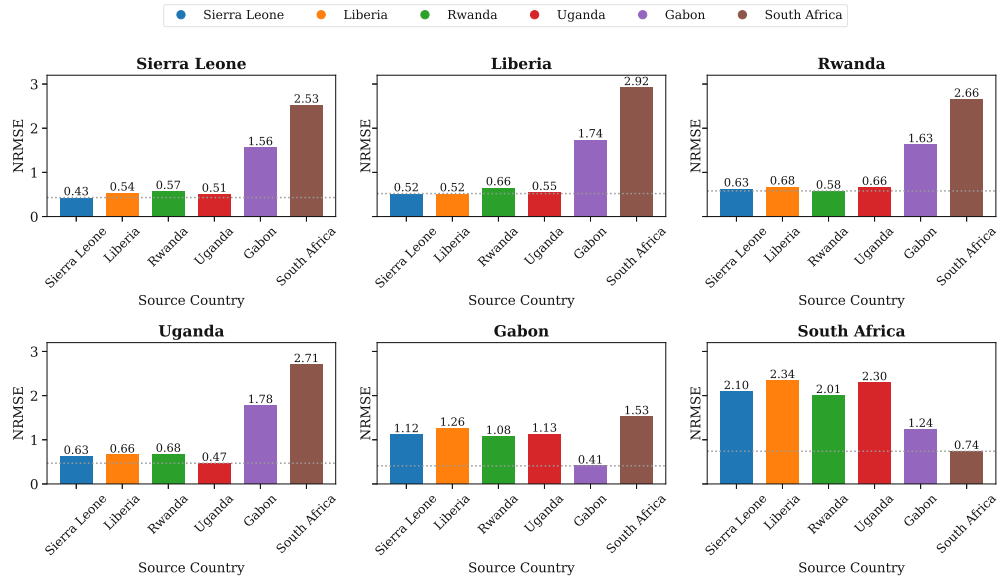


Figure 5.3: Cross-country model transfer performance relative to the within-country baseline. Each model is trained on the full source-country dataset and evaluated on the full target-country dataset. Results report NRMSE for predicted cluster means.

Table 5.2: Best performing source country per target country.

Target Country	Best Source Country
Sierra Leone	Uganda
Liberia	Sierra Leone
Rwanda	Sierra Leone
Uganda	Sierra Leone
Gabon	Rwanda
South Africa	Gabon

effectiveness across country pairs, assessing whether country similarity indices can predict these outcomes becomes important.

To evaluate the predictive value of similarity metrics, we compute the Pearson correlation between each metric’s country-to-country distance (or similarity) scores and the empirical model transfer performance. Specifically, for each similarity index, we construct the full pairwise similarity or distance matrix across countries. These values are then aligned with the corresponding empirical transfer results, measured as normalized root mean squared error (NRMSE) of model predictions in the target country. The Pearson correlation coefficient is used to quantify the strength of association: a high positive correlation indicates that increasing similarity between two countries is systematically associated with improved transfer performance. For indices where higher values indicate greater

similarity rather than distance, the correlation is sign-adjusted to ensure comparability across all metrics. This correlation thus provides a global measure of how well each similarity index reflects actual model transferability patterns.

In addition to the correlation analysis, we also report how often each similarity metric correctly identifies the empirically best source country for a given target. Here, the top-ranked source according to the metric is compared with the country yielding the lowest NRMSE in transfer. This count of correct selections provides complementary insight into whether a metric can guide the choice of a single best transfer partner, though it is sensitive to small performance differences and does not capture overall association patterns. Therefore, Pearson correlation is used as the primary evaluation criterion.

Table 5.3 presents the similarity metrics ordered by their Pearson correlation with transfer performance, including only those significant at the  $p < 0.01$  level. The third column reports, for each metric, how many of the six target countries were correctly matched with their best-performing source.

Table 5.3: Country similarity indices ranked by Pearson correlation ( $r$ ) with transfer performance. The third column reports how often each metric correctly identified the best source country out of six targets.

Country Similarity Index	Pearson $r$	Correct selections
Jones Country Similarity Index	0.78	2
Feature Importance Ranking Similarity(b - top 50)	0.78	3
GDP-Based Economic Similarity (a)	0.64	1
National Cultural Dimensions	0.59	0
Rural Population Density Distribution (a)	0.56	3
Polycentricity of Settlements (a)	0.55	4
Rural Population Density Distribution (d)	0.54	2
GDP-Based Economic Similarity (b)	0.51	1
Rural Population Density Distribution (b)	0.50	1

Several similarity indices demonstrate strong correlations with transfer performance, establishing their effectiveness for source country selection. The Jones Country Similarity Index and Feature Importance Ranking Similarity (top 50 features) both achieve the highest Pearson correlation coefficient of 0.78, indicating substantial predictive power for model transferability. The Jones Country Similarity Index represents a composite measure aggregating diverse socioeconomic, geographic, and demographic indicators across countries. In contrast, the feature importance-based metric quantifies the overlap between the top  $k$  most important features derived from models trained in each country pair.

Despite equivalent correlation performance, these metrics differ in practical applicability. The feature importance-based metric requires access to feature importance rankings from models trained on both source and target countries, which is not available in transfer

scenarios where target labels are unavailable. Furthermore, its performance is sensitive to the parameter  $k$ , with substantially weaker correlations observed for alternative values ( $k = 5, 10, 20, 100$ ). By contrast, the Jones index is readily applicable across settings. Figure 5.4 illustrates the relationship between the Jones Country Similarity Index and empirical model transfer performance.

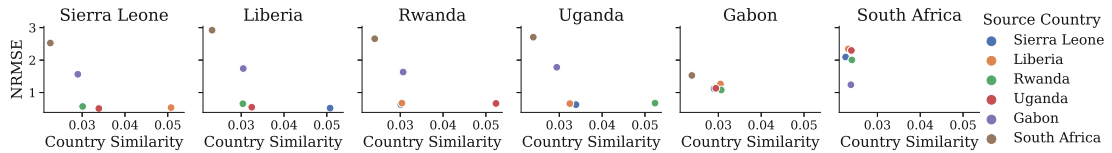


Figure 5.4: Scatter plots illustrating the association between the Jones Country Similarity Index and model performance. Each panel represents one target country, with colors indicating the corresponding source country.

The association between the Jones Country Similarity Index and model performance is clearly visible in Figure 5.4, although the relationship is not equally strong across all countries. For South Africa, for example, all source countries appear roughly equally similar according to the index, yet Gabon achieves distinctly better transfer performance. Another notable observation is that the model performance for Sierra Leone, Liberia, Uganda, and Rwanda clusters within a relatively narrow range.

Considering the secondary evaluation criterion, the number of cases in which the most similar country according to a metric coincides with the empirically best-performing source, further variability emerges. The Jones Country Similarity Index achieves two correct identifications, whereas for example the Polycentricity of settlements metric achieves four. However, in contexts where transfer performance differences are small, as observed for Sierra Leone, Liberia, Uganda, and Rwanda, minor measurement variations can alter rankings without substantial practical implications. Consequently, the Pearson correlation coefficient provides a more robust assessment of predictive validity.

### 5.2.2 Multi-Country Training for Model Transfer

The single-source transfer analysis revealed significant variability in performance, showing that even the empirically best single source often yields suboptimal results compared to within-country training and can perform poorly for certain targets (e.g., Gabon, South Africa). This motivates exploring whether training models on pooled datasets, constructed by incrementally adding data from the most successful individual source countries identified earlier, can enhance predictive performance on a target country. This approach tests the hypothesis that integrating information from several relevant contexts, despite increasing overall data diversity and potential dissimilarity, could allow the model to capture more generalizable features or relationships than is possible when relying on data from only one source nation.

Table 5.4: Normalized Root Mean Squared Error (NRMSE) for multi-source models across six target countries. Best performance for each country (i.e., lowest NRMSE) is highlighted in bold.

Target Country	1 Source	2 Sources	3 Sources	4 Sources	5 Sources
Sierra Leone	0.51	0.48	0.48	<b>0.47</b>	0.48
Uganda	0.63	0.60	<b>0.58</b>	0.59	0.58
Rwanda	0.63	0.62	0.62	<b>0.58</b>	0.65
Liberia	<b>0.52</b>	0.57	0.58	0.56	0.62
Gabon	1.08	1.09	0.99	0.92	<b>0.77</b>
South Africa	<b>1.24</b>	1.69	1.55	1.68	1.77

Multi-source training improved model performance in four out of six target countries, reducing the Normalized Root Mean Squared Error (NRMSE) toward the within-country training baseline. The most substantial improvement was observed for Gabon, where single-source transfer had previously failed (NRMSE > 1). By integrating data from all available source countries, the model achieved a notable reduction in error (NRMSE = 0.77).

Examining the wealth index distributions (see Figure 3.2) provides intuition for these results. Sierra Leone, Liberia, Uganda, and Rwanda have similar wealth distributions, clustering closely together. South Africa is an outlier with significantly higher wealth and small distributional overlap with these four countries. Gabon’s wealth distribution lies between South Africa and the clustered group.

This context explains why multi-source training fails to improve performance for South Africa: including countries even more dissimilar than Gabon introduces irrelevant patterns. For Gabon, however, integrating all countries proves optimal, as it captures socioeconomic patterns from both the poorer cluster and the wealthier South African context, effectively bridging their characteristics. In Liberia, multi-source training yields no improvement compared to single-source training. The near-identical wealth distributions of Liberia and Sierra Leone suggest that Sierra Leone alone as a source is sufficient, rendering additional sources disruptive. Notably, the reverse is not true: when Sierra Leone is the target, Liberia is not the optimal single source, indicating that distributional similarity does not always guarantee the best transfer performance.

These findings suggest that pooling data from related sources can improve generalization, but including overly dissimilar sources may result in negative transfer.

To better understand when in-country training should be preferred over model transfer, we compared the performance of within-country models trained on increasing amounts of data to the respective best multi-source transfer model. Figure 5.5 shows this comparison for the four countries found to cluster more closely together: Sierra Leone, Liberia, Uganda and Rwanda.

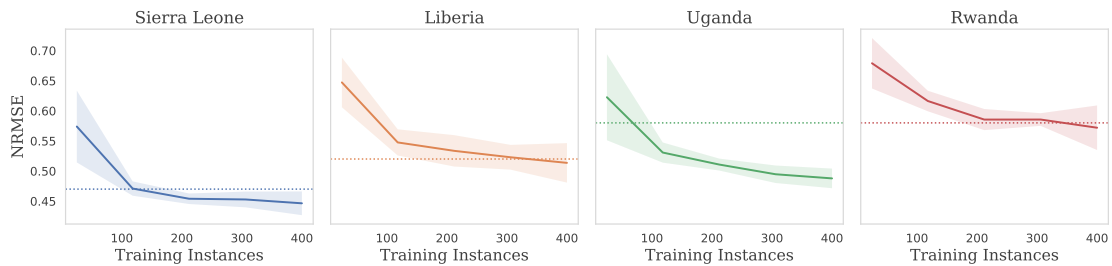


Figure 5.5: Mean NRMSE of the CatBoost regressor versus training set size for Sierra Leone, Liberia, Uganda and Rwanda. Solid lines show the average NRMSE over eight random splits for training set sizes of 25 to 400 instances, with shaded areas representing  $\pm 1$  standard deviation. Dotted horizontal lines indicate the best country-specific result from multi-source model transfer.

The results show that the amount of in-country data needed to surpass the model transfer baseline varies across countries. In Sierra Leone and Uganda, as few as 100 labeled instances are sufficient to outperform the best transfer model. In contrast, Liberia and Rwanda require approximately 350 training instances, almost the full size of their available datasets, to achieve similar performance.

This suggests that model transfer is most valuable when in-country data is very limited. As the number of labeled instances increases, in-country training quickly becomes preferable, particularly in settings where the target country’s data distribution is not easily captured by external models.

### 5.2.3 Comparison of Mitigation Strategies for Missing Features

In many low-income regions, certain types of data may be incomplete or entirely unavailable. This presents a practical challenge for model transfer, where models trained on data-rich source countries are applied to data-sparse target countries. This section systematically evaluates strategies to mitigate the impact of missing feature categories in the target country.

To simulate these scenarios, one feature category is removed at a time in the target country while the full feature set remains available in the source country. Two mitigation strategies are compared. The first is a retraining approach, where the model is trained on the source country using only the subset of features available in the target. The second is a feature reconstruction approach, where the model is trained on the full feature set in the source country and a separate model is used to reconstruct the missing features in the target country based on the remaining ones. This reconstruction model is also trained using data from the source country.

As a source country, the one previously identified as having the highest transfer performance is used. This analysis aims to assess which strategy is more effective in maintaining predictive accuracy when faced with partial data availability in the target setting.

Figure 5.6 summarizes the performance of different model transfer strategies under simulated missing data scenarios across all target countries.

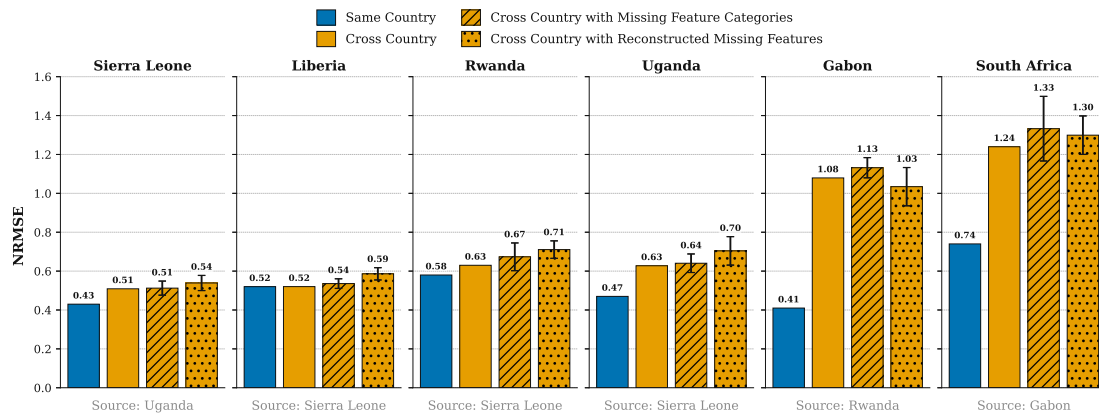


Figure 5.6: Performance comparison of model transfer strategies under simulated missing data across all countries. The bars represent: (1) the within-country baseline performance, (2) cross-country performance using the best source country with all features, (3) cross-country performance when retraining the model on a reduced feature subset, and (4) performance using feature reconstruction. Results are shown as the average NRMSE across scenarios where one of the four feature categories is removed in the target country. Error bars indicate the standard deviation across these missing-feature scenarios.

Two primary observations emerge from this analysis. Firstly, omitting a single feature category typically results in no or only minimal reductions in cross-country model performance. This suggests that the models are relatively robust to partial data loss. A likely explanation is the presence of sufficient redundancy across feature categories so information lost from one category can be compensated for by the remaining features. Secondly, the feature reconstruction strategy consistently reduces predictive accuracy compared to retraining models solely on available features, except in Gabon and South Africa, where model transfer had already proven ineffective. The diminished performance from reconstruction is likely due to inaccuracies in estimated features, which propagate errors into wealth index predictions. These findings suggest that model simplification through feature subset training represents a more robust strategy than feature imputation when handling missing categories in cross-country applications.

### Feature Reconstruction Performance on Source and Target Countries

To better understand the limitations of the feature reconstruction strategy, this section evaluates reconstruction accuracy as progressively more features are removed from the target dataset. Importantly, the features are removed in order of decreasing importance—starting with those ranked most important based on their contribution to model performance. For each target country, a feature reconstruction model is trained on

the corresponding source country, which is also the origin of the transfer model. This reconstruction model estimates the values of the removed features in the target country using the remaining available features. Figure 5.7 presents the average NRMSE for all reconstructed features, comparing performance in the source and target countries as a function of the number of removed features.

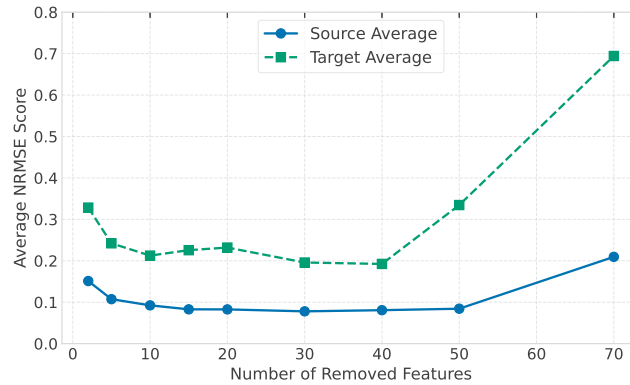


Figure 5.7: Average feature reconstruction performance across all six countries, measured by NRMSE. Blue circles indicate reconstruction error in the source country (training domain), and green squares indicate performance in the target country (application domain), shown across different numbers of removed features.

As expected, reconstruction accuracy is consistently higher in the source country, reflecting the domain shift that occurs when models are applied to a different country than the one they were trained on. Reconstruction error is relatively high when only a few features are removed, potentially suggesting these initial features contain unique information that is inherently difficult to predict from the remaining covariates. As the number of removed features increases substantially, the reconstruction error rises again, likely reflecting the diminishing amount of predictive information available in the remaining features to train an effective reconstruction model. This loss in reconstruction quality when applied across countries likely introduces additional noise, which in turn reduces predictive performance in the poverty estimation models.

### 5.3 Transfer Learning

The cross-country experiments revealed that direct model transfer is viable only when the source and target domains are highly similar. Even then, within-country models still set the performance ceiling. Given the scarcity of household survey data in certain regions, transfer learning presents a promising approach: it enables the use of abundant labeled data from resource-rich countries to pre-train a model, which can then be adapted to the limited labeled data available in a data-poor target country.

In theory, this strategy can (i) reduce the quantity of ground truth required in the target domain, (ii) mitigate negative transfer by fine-tuning to the target distribution, and (iii) result in models that are overall more robust to cross-country heterogeneity than models trained from scratch.

To test whether transfer learning can improve on the within-country model baseline, several techniques are evaluated.

**Multi-country.** In this setting, the training data from all available countries are pooled into a single dataset to train the model jointly. To account for systematic differences between countries, an additional categorical variable indicating country membership is appended to each observation. This approach allows the model to exploit the larger training sample while still distinguishing between country-specific contexts, thereby testing whether simple data pooling can improve transfer performance over the within-country baseline.

**Model Stacking.** All six single-source CatBoost models are first trained on their respective countries. Their predictions on the target country’s features are then stacked as additional covariates when fitting a new CatBoost model on the target’s training folds. Stacking thus supplies a high-level, cross-country prior while still allowing the final learner to specialise to local patterns.

**TransTab.** TransTab [WS22] is a deep learning–based adaptation of TabNet designed for transfer learning on tabular data. In this study, TransTab was trained on the pooled source data using supervised learning, followed by fine-tuning on the target country’s labeled data. The model leverages gated attention mechanisms to learn sparse, transferable feature representations, aiming to adapt effectively to the target distribution during fine-tuning.

**CORAL.** Correlation Alignment (CORAL) is a domain adaptation method that reduces distributional differences between source and target data by aligning their covariance structures [SFS16]. In this study, source features are linearly transformed to match the target covariance, thereby mitigating cross-country domain shifts and improving model transferability.

**TrAdaBoostR2.** Transfer AdaBoost [DYXY07] iteratively down-weights source instances that hurt performance on the target validation split while up-weighting informative target instances. Two variants were tested: (i) using only the empirically best single source identified earlier, and (ii) using the best sources from the multi-country experiment. The differences were marginal, the best results are shown.

**Feature Augmentation.** Feature Augmentation (FA) [DI07] expands the feature space to separate domain-general from domain-specific information. Each original feature is replicated into shared, source-specific, and target-specific versions, allowing the

model to learn which aspects transfer across countries and which remain context-dependent. In multi-country settings, this generalizes to multiple domain-specific copies plus a common feature set. FA thus enables standard learners to handle cross-country discrepancies while leveraging shared structure.

All methods were evaluated under identical conditions using 5-fold cross-validation and Bayesian hyperparameter optimization with 20 iterations to ensure comparability. For each method and target country, multiple experiments were conducted in which between one and five source countries were incrementally added as auxiliary training data. Source countries were ranked according to their performance in the previous model transfer experiments, and data were added following this order. The best-performing configuration for each method and country is reported. Table 5.5 summarizes the results, comparing all transfer learning techniques against the within-country CatBoost baseline.

Table 5.5: Best NRMSE achieved by each transfer learning method across the six target countries. Bold values indicate improvements over the within-country CatBoost baseline, and blue highlights mark the best performance per country.

Method	Sierra L.	Liberia	Uganda	Rwanda	Gabon	South A.
CatBoost Baseline	0.426	0.522	0.467	0.583	0.413	<b>0.738</b>
Multi-country	0.426	<b>0.515</b>	0.470	<b>0.575</b>	<b>0.408</b>	0.741
Model Stacking	0.432	0.525	0.472	0.583	0.420	0.756
TransTab	0.428	0.563	0.514	<b>0.577</b>	0.430	0.756
CORAL	0.433	<b>0.509</b>	0.479	<b>0.575</b>	<b>0.404</b>	0.739
TrAdaBoostR2	0.438	<b>0.518</b>	<b>0.465</b>	0.594	0.421	0.750
Feature Augm.	<b>0.424</b>	<b>0.507</b>	<b>0.466</b>	<b>0.570</b>	<b>0.404</b>	0.740

The Multicountry approach achieves modest improvements, particularly in Liberia, Rwanda, and Gabon, showing that pooling data from multiple countries combined with a country indicator variable can enhance predictive performance. Model Stacking does not yield systematic gains and often performs marginally worse, suggesting that simple ensembling of source predictions introduces noise rather than useful cross-country information.

TransTab likewise shows no consistent benefit in this application. While it matches the baseline in Rwanda, performance deteriorates in Liberia, Uganda, and Gabon, indicating that the additional model complexity does not necessarily translate into effective transfer when training data are limited.

TrAdaBoostR2 closely matches within-country performance, with its re-weighting scheme converging toward the target-only learner. CORAL achieves improved results in some settings, but the best-performing method overall is Feature Augmentation. It improves upon the baseline in five out of six countries and delivers the best results of all methods in four of them. Its explicit separation of shared and domain-specific features appears

particularly well-suited to addressing cross-country heterogeneity in poverty mapping data.

With respect to the number of source countries used, Feature Augmentation performed best with a single additional source for Liberia and Uganda, with two for Gabon, and with three for Sierra Leone, Rwanda, and South Africa. This suggests that adding more sources beyond a certain similarity threshold does not yield further improvement, an observation consistent with the earlier model transfer experiments.

Although the performance gains are moderate the results show the potential of transfer learning methods to enhance predictive accuracy beyond within-country training.

## 5.4 Transferability decision framework

To summarize the findings, a transferability decision framework is proposed to guide the development for high accuracy poverty maps depending on which data sources are available (Figure 5.8).

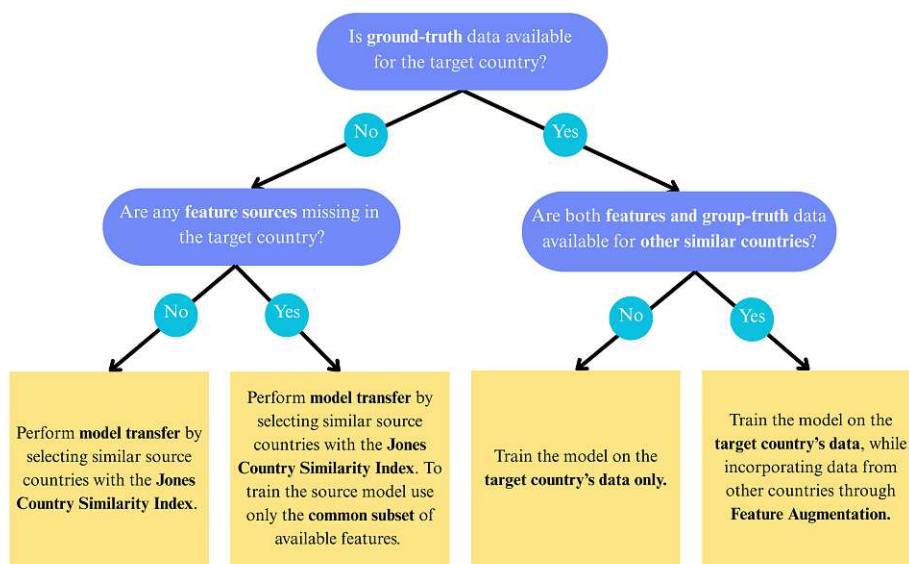


Figure 5.8: Decision framework to guide the development for high accuracy poverty maps depending on which data sources are available.

If no ground-truth data are available for the target country, model transfer becomes necessary. In this case, the Jones Country Similarity Index is used to identify the most

## 5. RESULTS

---

suitable source countries. When all feature sources are available in the target country, the source model can be transferred directly. If, however, some feature sources are missing, only the common subset of features across source and target is used for training.

If ground-truth data are available in the target country, models can in principle be trained directly. When both features and ground-truth data are also available for similar countries, Feature Augmentation offers an additional strategy: target-country data form the basis of model training, while information from other countries is incorporated at the preprocessing stage.

This framework integrates the main conclusions from the evaluation of similarity metrics and transfer strategies. It provides a systematic guide to select the most appropriate modeling strategy depending on the data environment of the target country.

# Conclusion

This chapter summarizes the main findings of the thesis, highlighting their contributions to the literature on poverty mapping with machine learning in data-scarce environments. The chapter concludes by outlining the study's limitations and identifying directions for future research.

## 6.1 Discussion and Contributions

This thesis explores machine learning-based poverty mapping in data-scarce environments. Addressing three research questions, the study examines the role of country similarity in model transfer, optimal approaches for handling incomplete data and the potential of transfer learning to enhance predictive performance beyond in-country training. The findings provide actionable insights for poverty mapping in resource-constrained settings.

To address the challenge of limited survey data, this study investigates the transferability of poverty mapping models across countries. While within-country models consistently achieve the highest predictive performance, cross-country transfers can yield comparable results, particularly between similar countries. However, transfer effectiveness varies substantially depending on the country pair, highlighting the need for informed source country selection.

**RQ1: By how much does Country Similarity affect the Performance of Model Transfer?** The results confirm that country similarity significantly influences the performance of model transfer in poverty mapping. Models transferred between countries with similar socio-economic and geographic characteristics, such as Sierra Leone, Liberia, Uganda, and Rwanda, exhibit only minor reductions in predictive accuracy compared to within-country models. In contrast, transfers involving countries with divergent characteristics, such as Gabon and South Africa, result in substantial performance degradation, often failing to outperform a baseline of predicting the mean ( $\text{NRMSE} > 1$ ).

Among the evaluated similarity metrics, the Jones Country Similarity Index was the most effective, showing a strong association with transfer performance (Pearson  $r=0.78$ ). Its practical advantage is that it is readily available and does not require additional data. These findings support the hypothesis that robust, easily deployable similarity metrics can provide guidance for source data selection in settings with limited ground-truth data.

The multi-country training experiments showed that pooling data from multiple similar sources generally improves transfer performance over single-source training likely by providing more diverse and representative training data. For instance, Gabon saw significant improvement when data from all available sources were included. However, the selection of source countries is critical. In South Africa, multi-source training degraded performance, as including dissimilar countries introduced irrelevant patterns. These findings underscore the importance of using similarity metrics to guide source country selection for multi-source training.

While not a substitute for up-to-date household surveys, which produce the most accurate poverty maps, model transfer can effectively bridge gaps in survey data availability, supporting targeted interventions in data-scarce regions.

### **RQ2: Optimal Approach for Generating Poverty Maps with Missing Data.**

To address the common challenge of incomplete data in low-resource settings, this study evaluates strategies for generating poverty maps when both ground-truth data and certain feature categories are unavailable in the target country. Two approaches were compared: (1) retraining a model on the source country using only the features available in the target country, and (2) reconstructing missing features in the target country using a model trained on the source country before applying the transfer model. These experiments simulated real-world scenarios by systematically removing one feature category at a time in the target country while retaining the full feature set in the source country.

The results show that retraining on the available feature subset consistently outperforms feature reconstruction, as the latter appears to introduce additional noise that reduces predictive accuracy. Across most target countries and simulated missing feature categories, retraining the model on the available subset maintained performance close to the original cross-country transfer using the full feature set. In practical applications, where different data sources may be missing across countries due to varying levels of technology adoption or data availability, this finding is particularly important: it shows that such gaps do not necessarily undermine the effectiveness of poverty mapping models. The observed resilience potentially arises from information redundancy among geospatial feature types, allowing the models to compensate for missing categories.

### **RQ3: To What Extent Can Transfer Learning Enhance Predictive Performance?**

The results show that transfer learning can provide measurable, though generally moderate, improvements over within-country baselines. Among the tested methods, Feature Augmentation proved most effective, outperforming the baseline in five of six target countries and achieving the overall best results in four. CORAL and Multi-country

training also yielded gains in certain cases, while Model Stacking and TransTab offered no benefit.

Feature Augmentation’s explicit separation of shared and domain-specific features makes it particularly well suited to cross-country heterogeneity. By contrast, more complex neural network-based approaches such as TransTab did not lead to better performance, suggesting that additional model complexity is not advantageous in this setting. Representation-transfer methods like TransTab are likely more appropriate when complex feature interactions are central and when substantially larger and more diverse training corpora are available.

An important insight is that the effectiveness of transfer learning depends not only on the chosen method but also on the number and similarity of included source countries. Feature Augmentation, for example, performed best with one or two closely related sources, while adding more distant countries did not improve performance. This finding aligns with earlier transfer experiments and underscores the risk of negative transfer when dissimilar contexts are included.

These results add to the existing literature, where most studies pool all available countries as training data [LB22, CFCB22]. Such an approach typically improves performance over within-country training in only a small fraction of countries. By contrast, this framework achieves improvements in the majority of cases.

Finally, it is important to note that the countries included in this study were deliberately selected for their diversity in order to evaluate the role of country similarity. This choice runs counter to the optimal conditions for transfer learning, where having many similar source countries to the target would be most beneficial. A dataset constructed with this goal in mind would likely increase the magnitude of improvements achievable with transfer learning.

The empirical findings are synthesized into a practical framework that recommends: (i) similarity-guided model transfer when labels are unavailable, (ii) subset retraining when features are missing, and (iii) Feature Augmentation when comparable external data can be leveraged.

## 6.2 Limitations and Future Research

Several limitations remain that open avenues for future research. These include:

1. The analysis in this thesis was limited to six Sub-Saharan African countries, which may not fully capture the diversity of global poverty contexts. Expanding the scope to include additional countries, particularly from other regions with different socio-economic, demographic, and environmental conditions, would strengthen the robustness and generalizability of the findings. Future work could therefore test whether the observed patterns, for example the best performing country similarity index, hold across a broader range of contexts.

2. A key finding of this thesis is that indiscriminately pooling data from multiple source countries can be counterproductive, as negative transfer can degrade performance. This was observed in both model transfer and in the transfer learning experiments. This finding challenges the common practice in other studies of using all available countries as source data. With a larger sample of countries, future work could systematically investigate this phenomenon. A compelling research direction would be to determine if a specific similarity threshold exists, beyond which adding another country's data is more likely to harm than help performance.
3. As mentioned earlier, assembling a larger corpus of similar countries for a given target may further improve the effectiveness of transfer learning. Future research could test this systematically by comparing diverse versus similarity-focused country samples. Another promising direction is to assess whether transfer learning contributes to fairness, for example by reducing performance gaps between rural and urban areas. A practical limitation of this study is that the cross-country experiments were highly compute-intensive, which restricted the number of feasible train–test iterations. While running additional iterations could yield more robust and reliable performance estimates, this comes at considerable computational and time costs.
4. Another limitation is the prediction of cluster-level wealth standard deviations. While this measure could provide valuable insights into local inequality and vulnerability, the models did not achieve strong performance, limiting its practical utility in the current setup.
5. The models may be vulnerable to rapid economic or environmental shocks, such as pandemics or climate-related events, which can disrupt the relationship between features like night-lights and income. This study did not account for the temporal dimension and prior research suggests that capturing poverty dynamics over time remains challenging [KZ20]. Future work could explore time-adaptive models to improve resilience to such disruptions.
6. While this study focused on tabular models, future work could explore integrating tabular approaches with satellite image based models. Satellite imagery provides complementary information to structured metadata, and prior research has shown that hybrid models combining both modalities often achieve superior predictive performance. Extending the methods developed here to such multimodal settings may therefore represent a promising avenue for further improvement.

# Übersicht verwendeter Hilfsmittel

In the course of writing this thesis, generative AI models and tools were employed as writing aids and for code generation. These tools were used to support the writing process by suggesting improvements for formulation and clarity.

No text in this thesis has been included without substantial modification and refinement by me. The final manuscript reflects my independent intellectual work.

The following generative AI tools were used:

- **Chat-based models:** OpenAI ChatGPT (versions 4.5, 4o, and o3)
- **Code generation models:** Cursor agent, powered by Claude 3.7



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar  
The approved original version of this thesis is available in print at TU Wien Bibliothek.

# List of Figures

1.1	Inferred poverty maps for Sierra Leone (left) and Uganda (right), with color scales representing estimated values of the International Wealth Index [ENKK23]. . . . .	3
1.2	Overview of the fundamental machine learning setup used to obtain poverty maps. After Espin-Noboa et al. (2023) [ENKK23]. . . . .	4
1.3	This figure presents a scenario where a poverty map needs to be created for Uganda (target country), despite lacking a complete feature layer and ground-truth data. In this case, model transfer offers a solution by leveraging pre-trained models from similar countries, identified through a country similarity index. The most similar country, highlighted in orange, represents a suitable source for model transfer. If ground-truth data were available in Uganda, transfer learning could be used by loading the model from the source country and continuing training with data from Uganda. . . . .	5
3.1	Geographic locations of the Sub-Saharan African countries analyzed: Sierra Leone, Uganda, Liberia, Gabon, South Africa, and Rwanda. . . . .	18
3.2	Distribution of cluster-level mean IWI for six countries. Curves are Gaussian kernel-density estimates, so the area under each curve equals 1. . . . .	19
5.1	Feature importance distribution for Uganda and Sierra Leone. . . . .	36
5.2	Overall sum and average importance of feature categories over all countries. . . . .	37
5.3	Cross-country model transfer performance relative to the within-country baseline. Each model is trained on the full source-country dataset and evaluated on the full target-country dataset. Results report NRMSE for predicted cluster means. . . . .	38
5.4	Scatter plots illustrating the association between the Jones Country Similarity Index and model performance. Each panel represents one target country, with colors indicating the corresponding source country. . . . .	40
5.5	Mean NRMSE of the CatBoost regressor versus training set size for Sierra Leone, Liberia, Uganda and Rwanda. Solid lines show the average NRMSE over eight random splits for training set sizes of 25 to 400 instances, with shaded areas representing $\pm 1$ standard deviation. Dotted horizontal lines indicate the best country-specific result from multi-source model transfer. . . . .	42
		55

5.6	Performance comparison of model transfer strategies under simulated missing data across all countries. The bars represent: (1) the within-country baseline performance, (2) cross-country performance using the best source country with all features, (3) cross-country performance when retraining the model on a reduced feature subset, and (4) performance using feature reconstruction. Results are shown as the average NRMSE across scenarios where one of the four feature categories is removed in the target country. Error bars indicate the standard deviation across these missing-feature scenarios. . . . .	43
5.7	Average feature reconstruction performance across all six countries, measured by NRMSE. Blue circles indicate reconstruction error in the source country (training domain), and green squares indicate performance in the target country (application domain), shown across different numbers of removed features. . . . .	44
5.8	Decision framework to guide the development for high accuracy poverty maps depending on which data sources are available. . . . .	47

## List of Tables

3.1	Number of clusters per country . . . . .	19
5.1	Model results for wealth index cluster mean and standard deviation predictions.	36
5.2	Best performing source country per target country. . . . .	38
5.3	Country similarity indices ranked by Pearson correlation ( $r$ ) with transfer performance. The third column reports how often each metric correctly identified the best source country out of six targets. . . . .	39
5.4	Normalized Root Mean Squared Error (NRMSE) for multi-source models across six target countries. Best performance for each country (i.e., lowest NRMSE) is highlighted in bold. . . . .	41
5.5	Best NRMSE achieved by each transfer learning method across the six target countries. Bold values indicate improvements over the within-country CatBoost baseline, and blue highlights mark the best performance per country.	46

# Bibliography

- [AP21] Sercan Ö. Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6679–6687, 2021.
- [ARB23] Emily Aiken, Esther Rolf, and Joshua Blumenstock. Fairness and representation in satellite-based poverty maps: Evidence of urban–rural disparities and their impacts on downstream policy. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI 2023)*, 2023.
- [BMV<sup>+</sup>23] H. E. Beck, T. R. McVicar, N. Vergopolan, M. Pan, and E. F. Wood. High-resolution (1 km) köppen-geiger maps for 1901–2099 based on constrained cmip6 projections. *Scientific Data*, 10:724, 2023.
- [BNAP24] Sina Bagheri Nezhad, Ameeta Agrawal, and Rhitabrat Pokharel. Beyond data quantity: Key factors driving performance in multilingual language models. *arXiv preprint arXiv:2412.12500*, 2024.
- [CFCB22] Guanghua Chi, Han Fang, Sourav Chatterjee, and Joshua E. Blumenstock. Microestimates of wealth for all low- and middle-income countries. *Proceedings of the National Academy of Sciences*, 119(3):e2113658119, 2022.
- [DI07] Hal Daumé III. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 256–263, Prague, Czech Republic, June 2007. Association for Computational Linguistics.
- [DYXY07] Wenyuan Dai, Qiang Yang, Gui-Rong Xue, and Yong Yu. Boosting for transfer learning. In *Proceedings of the 24th international conference on Machine learning*, pages 193–200. ACM, 2007.
- [ELL03] Chris Elbers, Jean O. Lanjouw, and Peter Lanjouw. Micro-level estimation of poverty and inequality. *Econometrica*, 71(1):355–364, 2003.
- [ENKK23] Lisette Espín-Noboa, János Kertész, and Márton Karsai. Interpreting wealth distribution via poverty map inference using multimodal data. In *Proceedings of the ACM Web Conference 2023, WWW '23*, page 4029–4040. ACM, April 2023.

- [GPS24] Josh Gardner, Juan C. Perdomo, and Ludwig Schmidt. Large scale transfer learning for tabular data via language modeling. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024.
- [GRKB21] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, pages 18932–18943, 2021.
- [HKCK20] Xin Huang, Ashish Khetan, Milan Cvitkovic, and Zohar Karnin. Tabtransformer: Tabular data modeling using contextual embeddings. *arXiv preprint arXiv:2012.06678*, 2020.
- [Hof80] Geert H. Hofstede. *Culture’s Consequences: International Differences in Work-Related Values*. Sage Publications, Beverly Hills, CA, 1980.
- [HS02] Norbert Henninger and Mathilde Snel. *Where are the Poor? Experiences with the Development and Use of Poverty Maps*. World Resources Institute and UNEP/GRID-Arendal, Washington, DC and Arendal, Norway, 2002.
- [JBX<sup>+</sup>16] Neal Jean, Marshall Burke, Michael Xie, W. Matthew Alampay Davis, David B. Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016.
- [Jon25] Jeff M. Jones. Country similarity index. <https://objectivelists.com/country-similarity-index/>, 2025. Objective Lists. First published 2020; updated 2025. Accessed 27 September 2025.
- [KKEN24] Márton Karsai, János Kertész, and Lisette Espín-Noboa. A comparative analysis of wealth index predictions in africa between three multi-source inference models. *arXiv preprint arXiv:2408.01631*, 2024.
- [KZ20] Lukas Kondmann and Xiao Xiang Zhu. Measuring changes in poverty with deep learning and satellite images. In *Proceedings of the ICLR 2020 Workshop on Practical Machine Learning for Developing Countries*, 2020.
- [LB22] Kamwo Lee and Jeanine Braithwaite. High-resolution poverty maps in sub-saharan africa. *World Development*, 159:106028, 2022.
- [LCS<sup>+</sup>22] Roman Levin, Valeriia Cherepanova, Avi Schwarzschild, Arpit Bansal, C. Bayan Bruss, Tom Goldstein, Andrew Gordon Wilson, and Micah Goldblum. Transfer learning with deep tabular models. In *Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS 2022) — Workshop on Table Representation Learning*, 2022. Published as conference paper at ICLR 2023.
- [MD24] Robert Marty and Alice Duhaut. Global poverty estimation using private and public sector big data sources. *Scientific Reports*, 14(1):3160, 2024.

- [MTB13] Manuel Magombeyi, Akpofure E. Taigbenu, and Jennie Barron. Rural poverty and food insecurity mapping at district level for improved agricultural water management in the limpopo river basin. Technical Report CPWF Research for Development (R4D) Series 6, CGIAR Challenge Program on Water and Food (CPWF), Colombo, Sri Lanka, 2013.
- [NHAJ24] Asmik Nalmpatian, Christian Heumann, Levent Alkaya, and William Jackson. Transfer learning for mortality risk: A case study on the united kingdom. *medRxiv preprint*, 2024.
- [PGV<sup>+</sup>18] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: Unbiased boosting with categorical features. In *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS)*, Montréal, Canada, 2018.
- [PJ17] Nitya Pokhriyal and David C. Jacques. Combining disparate data sources for improved poverty prediction and mapping. *Proceedings of the National Academy of Sciences of the United States of America*, 114(46):E9783–E9792, Nov 2017. Epub 2017 Oct 31.
- [PWP23] Salwa Rizqina Putri, Arie Wahyu Wijayanto, and Setia Pramana. Multi-source satellite imagery and point of interest data for poverty mapping in east java, indonesia: Machine learning and deep learning approaches. *Remote Sensing Applications: Society and Environment*, 29:100889, 2023.
- [RLHK10] Clionadh Raleigh, Andrew Linke, Håvard Hegre, and Joakim Karlsen. Introducing acled: An armed conflict location and event dataset. *Journal of Peace Research*, 47(5):651–660, 2010.
- [SB22] Isaac S Smythe and Joshua E Blumenstock. Geographic microtargeting of social assistance with high-resolution poverty maps. *Proceedings of the National Academy of Sciences*, 119(32):e2120025119, 2022.
- [SCD<sup>+</sup>22] Mehdi Shojaie, Mercedes Cabrerizo, Steven T. DeKosky, David E. Vaillancourt, David Loewenstein, Ranjan Duara, and Malek Adjouadi. A transfer learning approach based on gradient boosting machine for diagnosis of alzheimer’s disease. *Frontiers in Aging Neuroscience*, 14, 2022.
- [SFM<sup>+</sup>18] Linda See, Steffen Fritz, Inian Moorthy, Olha Danylo, Michiel Van Dijk, and Barbara Ryan. *Using remote sensing and geospatial information for sustainable development*, pages 172–198. Brookings Institution Press, 2018.
- [SFS16] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*, pages 2058–2065, Phoenix, Arizona, USA, 2016. AAAI Press. Introduces CORAL (Correlation Alignment).

- [She01] Oded Shenkar. Cultural distance revisited: Towards a more rigorous conceptualization and measurement of cultural differences. *Journal of International Business Studies*, 32(3):519–535, 2001.
- [SLW<sup>+</sup>22] Yiheng Sun, Tian Lu, Cong Wang, Yuan Li, Huaiyu Fu, Jingran Dong, and Yunjie Xu. Transboost: A boosting-tree kernel transfer learning algorithm for improving financial inclusion. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI 2022)*, pages 12181–12190, 2022.
- [The22] The DHS Program. The demographic and health surveys (dhs) program. <https://dhsprogram.com/>, 2022.
- [WCC<sup>+</sup>25] Haoxiang Wang, Xiaoping Che, Enyao Chang, Chenxin Qu, Ganghua Zhang, Zihan Zhou, Zhenlin Wei, Gengyu Lyu, and Pengfei Li. Similarity based city data transfer framework in urban digitization. *Scientific Reports*, 15(1):10776, 2025.
- [WS22] Zifeng Wang and Jimeng Sun. Transtab: Learning transferable tabular transformers across tables. In *Advances in Neural Information Processing Systems (NeurIPS 2022)*, 2022.
- [YD10] Yi Yao and Gianfranco Doretto. Boosting for transfer learning with multiple sources. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1855–1862. IEEE, 2010.
- [YPD<sup>+</sup>20] Christopher Yeh, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon, and Marshall Burke. Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nature Communications*, 11(1):2583, 2020.
- [ZSE<sup>+</sup>23] Bingzhao Zhu, Xingjian Shi, Nick Erickson, Mu Li, George Karypis, and Mahsa Shoaran. Xtab: Cross-table pretraining for tabular transformers. In *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)*, 2023.