CHALLENGES OF INFERRING HIGH-RESOLUTION POVERTY MAPS WITH MULTIMODAL DATA

Lisette Espín-Noboa[1,2], **János Kertész**[1,2], **and Márton Karsai**[1,3]
[1] Central European University, Vienna, Austria
[2] Complexity Science Hub, Vienna, Austria
[3] Rényi Institute of Mathematics, Budapest, Hungary
{EspinL, KerteszJ, KarsaiM}@ceu.edu

ABSTRACT

Poverty maps—spatial representations of economic wealth—are essential tools for governments and NGOs to adequately allocate infrastructure and services in places in need. They also help to better understand social phenomena such as human mobility and segregation, and environmental problems induced by urbanization. Traditionally, such maps are inferred from Census and survey data, which are expensive and collected occasionally; thus, they commonly provide outdated and low-resolution socioeconomic information, especially in developing countries. Remotely sensed data combined with advanced machine learning methods provided a recent breakthrough in poverty map inference. However, these models are not optimized to produce accountable results that guarantee accurate predictions to all sub-populations including the rich and the poor or the urban and rural divide. In this paper, we touch upon the opportunities that multimodal data can offer to solve these issues, as well as the challenges of working with noisy, biased and sparse datasets for predicting high-resolution poverty maps in Sierra Leone.

1 INTRODUCTION AND RELATED WORK

The first Sustainable Development Goal set by the United Nations is to eradicate poverty by 2030 (United Nations, 2020). Although fewer people were living in extreme poverty around the world by 2018, the decline in poverty rates has slowed down ever since. This stagnation was partly due to the COVID-19 pandemic, but the ongoing impacts of conflicts and climate catastrophes set further barriers for progress in this direction (World Bank, 2018). Traditional data collection techniques, such as Census or survey methods, fail to follow the effects of such rapid changes, therefore new data collection techniques are required to capture the aftermath of unexpected global-impact events. The identification of places-in-need requires rapid, flexible and precise inference to inform the adequate allocation of resources, which are often misplaced due to coarse-grained and out-dated statistics provided by Census and survey data (Wisner et al., 2014).

The fast penetration of mobile phones worldwide (Adam & Minges, 2018; AG, 2017) set a precedent for collection and use of big data for social good. Mobile usage data like call detailed records (CDR) or mobile airtime payment transactions have been used to infer several socioeconomic aspects of people (Dong et al., 2014; Gao et al., 2019; Cruz et al., 2021), which in turn were applied to map socioeconomic effects on the structure and dynamics of the underlying social network (Eagle et al., 2010; Leo et al., 2016; 2018). One main limitation of CDRs, however, is that mobile phone data is proprietary and access is often granted via partnerships or purchase. On the contrary, the emergence of Web 2.0 technologies has opened new avenues for collecting and sharing online annotations or digital traces. In addition, open data initiatives have strengthen collaborations between industry, academia and the public sector (Meta, 2022), which makes it easier to study social, economic and environmental issues through the estimation of socioeconomic indicators from online data (Llorente et al., 2015; Aletras & Chamberlain, 2018; Abitbol et al., 2018; Abitbol & Morales, 2021).

In recent years, satellite imagery has attracted great attention as a means of inferring high-resolution poverty maps. As a first attempt, nightlight intensity of places has been shown to be a good estimator of economic activity (Ghosh et al., 2010), especially for non-extreme-poor areas (Pinkovskiy &

Table 1: **Location data summary**. Cluster locations are collected from the Demographics and Health Survey program (DHS). Populated places are extracted from OpenStreetMap (OSM), and they are rural if type $\in$ {village, hamlet and isolated dwellings}.

TYPE OF PLACE	COUNTS (%)
DHS clusters [2016, 2019]	**893 [336, 557]**
- urban	308 (34%)
- rural	585 (66%)
OSM Populated Places [2021]	**9568**
- urban	64 (0.70%)
- rural	9504 (99.3%)

Sala-i Martin, 2016). Moreover, they can even capture household consumption of the extreme poor when combined with daylight satellite images (Jean et al., 2016). Other approaches leverage the visual qualities of street-level imagery and daylight satellite images to detect objects (e.g., vehicles, infrastructure, terrain) and use them as proxies of wealth of neighborhoods (Abitbol & Karsai, 2020; Ayush et al., 2021; AyushBurak Uzkent et al., 2021). Since any geo-spatial dataset can be used to build poverty maps via the spatial correlations of socioeconomic indicators, advanced deep learning methods trained on multiple features (e.g., satellite images together with population density and demographic data of people) provide scalable and better predictions of wealth at high-resolution for low- and middle-income countries around the globe (Lee & Braithwaite, 2020; Chi et al., 2022).

While these methods represent a great advance towards scalable, time-variant and fine-grained poverty maps, they commonly optimize performance over representativeness, and strongly depend on the availability of all data sources. As a result, they might perform poorly in countries with scarce data and fall short in accuracy guarantees necessary for policy makers. In addition, most of these methods use machine learning algorithms, with good performance conditional to specific data engineering and parametrization choices. This limits their generalization and transferability potential, especially in countries in emergency undergoing rapid demographic and environmental changes. Although the parametrization of these models are well documented, their robustness against data scarcity and time is usually not evaluated.

We build towards this goal in this preliminary study, where we focus on Sierra Leone, a country characterized by extreme poverty (World Bank, 2018; United Nations, 2020), and propose regression models to predict wealth at $\approx$ 10K populated places from multiple online data sources. Using the last two household surveys conducted in 2016 and 2019 in the Demographics and Health Program (DHS) (The DHS Program, 2022a; Statistics Sierra Leone & ICF, 2016; 2020), we computed the international wealth index (IWI) (Smits & Steendijk, 2015) of 893 localized population clusters in the country. In addition, we extracted 172 metadata- and 784 image-features for each place using openly available data from OpenStreetMap, OpenCelliD, Facebook Marketing API, Facebook Data for Good, Google Earth Engine, and Google Maps Static API. We evaluate these sources independently and in combination to highlight their importance and interchangeability during the prediction. Our analysis sheds light on how to use open data to reliably infer high-resolution poverty maps. We made our code openly available (Espín-Noboa, 2022).

2 Data collection

Ground-truth wealth. DHS survey data (The DHS Program, 2022a) is commonly available at a household level and can be aggregated into clusters covering between 4 and 44 households in Sierra Leone. This granularity allows us to use it as ground-truth (GT) data to train and test machine learning models that predict the wealth (IWI) of places from the visual, infrastructural, or demographic features of their surrounding areas. Table 1 shows some statistics of the 893 DHS clusters used in this study. The majority of these clusters (66%) belong to rural areas, and their IWI scores vary from 4.2 to 72.7 with a mean of 26.2 and standard deviation of 15.6 [1].

[1] The International Wealth Index (IWI) (Smits & Steendijk, 2015) is the first comparable asset based wealth index covering the complete developing world and it ranges from 0 (extreme poor) to 100 (extreme rich).

Target places. In order to infer a high-resolution poverty map, we extracted other 9568 populated places. These locations were identified from OpenStreetMap (2022; 2021)(Mocnik et al., 2018) as settlements representing cities, towns, villages, hamlets or isolated dwellings. Similar to the GT data, the majority of these places (99.3%) are in rural areas (see Table 1).

Features. For each location we generated 172 metadata-features using six data sources. To extract infrastructure details of a given place (e.g., schools, banks, roads, intersections, and cell towers) we leveraged crowd-sourced data from two well known platforms, OpenStreetMap (2022) and OpenCelliD (Unwired Labs, 2021). We extracted population and movement counts from Facebook data for Good (Meta, 2022) and collected audience reach estimates (number of monthly active users of given demographics) via the Facebook Marketing API (Meta, 2022). Nightlight-intensity related features were extracted from the Google Earth Engine (Google, 2022). We generated additional 784 image-features using a convolutional neural network (CNN) on daylight satellite images downloaded from Google Maps Static API (Google, 2022)[2] (more details in Section 4.1).

3 CHALLENGES

Noisy ground-truth locations. Geo-located ground-truth data is often anonymized by adding noise to the location to preserve the confidentiality of survey respondents (The DHS Program, 2022b). While this is fundamental for ethical reasons, it induces uncertainty in the predictions. Previous work addressed this issue by either covering bigger areas around each cluster (Chi et al., 2022), or re-arranging locations in an iterative way while adjusting their wealth too (Lee & Braithwaite, 2020). While these options are plausible, it is unclear by how much they affected the prediction. We propose to re-arrange the cluster locations only in rural areas, which are more likely to be surrounded by non-populated places. We simply move the noisy location of a cluster to the location of the closest populated place—obtained from OpenStreetMap (2021)—without altering its wealth index. In case multiple clusters were assigned to the same populated place, we prioritized the cluster which had fewer other potential matches. We repeated this iteratively until all rural clusters were re-arranged. All displacements lie within the maximum boundaries of noise added to the original cluster location, i.e., between 0 and 5 Km with 1% of clusters displaced by 0 to 10 Km (The DHS Program, 2022b).

Feature importance. When collecting features from multiple sources, it is important to understand how useful each data source is at predicting the target variable. This allows to get rid of unimportant features, or focus on the best ones when resources are limited and a full model cannot be trained. We address this issue by comparing the performance of each data source alone and in combination.

Class imbalance. As shown in Table 1, the GT data is very skewed both in terms of wealth and urban/rural places. These imbalances need to be taken into account during the modeling phase to avoid misleading results and favoring the majority class. In this preliminary study, we show the recall of our model at the intersection of these two dimensions.

Sample size and time mismatch across data sources. When using multiple data sources another possible complication comes from the different times of recording of the different datasets. Here, as GT data we used the last two available DHS surveys, which date back to 2016 and 2019. This decision was not arbitrary. First, the performance of deep neural networks often improves with larger amounts of training data. Second, we needed closest-to-date GT data to match the features collected in 2021. However, the 2019 DHS data provided only 557 clusters, thus we opted to combine them with the 336 clusters recorded in 2016 (see Table 1). This provided us a larger dataset for modeling at the cost of matching old indicators of wealth with newer data features. We address these issues by studying the effects of data augmentation and data recency on our models.

4 EXPERIMENTS

4.1 MODELS

Metadata (XGB). We built an XGBoost regressor model (Chen & Guestrin, 2016) to predict the IWI score of DHS clusters. First, we extracted all 172 features for each cluster. Second, we partitioned

[2]Most datasets are publicly available except *movement maps* (available only for nonprofits and researchers upon request) and *satellite images* (which can be reached through Google Academic Research Grants).

Table 2: **Performance by GT location:** Re-assignment of cluster locations to the closest populated place improves the XGB model.

CONFIGURATION	XGB	
	R^2	**MSE**
Noisy	0.73	64.62
Re-arranged	**0.83**	**41.86**

Table 3: **Performance by augmentation:** The CNN model improves when the images in the training sample are augmented.

AUGMENTATION	CNN	
	R^2	**MSE**
None	0.72	68.72
Offline (18 techniques)	**0.74**	**64.93**

Table 4: **Performance by feature source:** Rows represent the seven data sources of interest: Facebook Marketing API, OpenStreetMap, Google Maps Static API, OpenCelliD, Google Earth Engine, Facebook Population, and Facebook Movement (between tiles prior COVID-19). For each data source we extracted a set of features which were used to build one CNN and six XGB models.

SOURCE	FEATURES (#)	Model	R^2	MSE
FB-MA	Audience reach (37)	XGB	0.47	130.3
OSM	Infrastructure (54)	XGB	0.72	68.88
G-MS	Daylight satellite images (784)	CNN	0.74	64.94
OCI	Cell towers (9)	XGB	0.77	55.81
G-EE	Nightlight intensity (36)	XGB	0.79	52.47
FB-P	Population density (9)	XGB	0.81	47.59
FB-MV	Movement maps (27)	XGB	0.81	47.24
All	(956)	CNN+XGB	0.82	43.53

the data into train (80%) and test sets (20%). We further used the train set for a 4-fold cross-validation, and tuned 13 hyper-parameters [3] via Random Search on 200 combinations.

Images (CNN). We train a second model that learns to predict IWI scores using daylight satellite images. This model is based on a 22-layer linear regression Convolutional Neural Network (CNN) architecture whose second last layer returns a 784-feature vector that is used to predict the final layer with a linear activation. We use the same data splits as before, and tuned 4 hyper-parameters [4] via Random Search on 200 combinations. Additionally, we applied 18 offline augmentation techniques [5] to 50% of the images in the training set to diversify our sample and reduce the risk of overfitting.

CNN+XGB: We feed the 784 features from the the second-last layer of the CNN to the XGB model to verify whether the metadata- and image-features together produce the highest performance.

4.2 RESULTS

Using the best cross-validated model (with lowest mean root square error) we predicted the IWI of DHS clusters in the test set using the XGB and CNN models separately. We measure performance using R^2 and mean-square-errors (MSE)[6] The higher the R^2 value (or lower MSE), the better performance. Note that, for the purpose of comparison, the split of the data into train, validation and test sets is the same across models. Also, all models carried out their own hyper-parameter tuning.

Re-arranging ground-truth locations (XGB). Table 2 shows the performance of our XGB model using the noisy cluster locations and the proposed re-arrangement of locations. We see that performance considerably improves when the rural clusters are moved to the closest rural populated place. We used these re-arranged GT locations as the base for the following experiments.

Data augmentation (CNN). In Table 3 we see that the CNN model with augmented data outperforms by 2.8% the CNN model without data augmentation.

[3] n_estimators, max_depth, learning_rate, booster, gamma, min_child_weight, max_delta_step, subsample, colsample_bytree, colsample_bylevel, colsample_bynode, reg_lambda and tree_method.

[4] learning_rate, optimizer, dropout and batch_size.

[5] flips, rotations, crops, rescaling, Guassian noise, brightness, darkness, and erasing.

[6] R^2 represents the proportion of the variance for the IWI score being explained by the independent variables. MSE refers to the average of the squares of the errors (between predicted and true values).

Table 5: **Recall of SES inference:** The CNN model performs best in poor areas, while the XGB model performs best in the rural-poor and urban-middle classes.

SES	**XGB**			**CNN**		
	ALL	**RURAL**	**URBAN**	**ALL**	**RURAL**	**URBAN**
Poor	94%	97%	0%	98%	98%	100%
Lower middle	80%	20%	93%	67%	40%	73%
Upper middle	75%	0%	79%	55%	0%	58%

Table 6: **Performance by time configuration:** Training and testing on the most recent survey gives the best performance followed by the combined model, which uses the last two surveys in both sets.

CONFIGURATION	**XGB**		**CNN**	
Train → Test	$\mathbf{R^2}$	**MSE**	$\mathbf{R^2}$	**MSE**
Combined → Combined	0.83	41.86	0.74	64.94
Past (2016) → Future (2019)	0.77	57.14	0.70	75.87
2016 → 2016 (5 years to 2021)	0.78	53.94	0.61	95.07
2019 → 2019 (2 years to 2021)	**0.85**	**36.27**	**0.76**	**61.82**

Feature source. We see in Table 4 that while most of the features can provide relatively good predictive performance, the mobility and population data from Facebook have the best features to predict IWI scores. Moreover, while all features provide better performance ($R^2 = 0.82$) than individually ($R^2 \leq 0.81$), they are slightly worse than the metadata-only features ($R^2 = 0.83$).

Sub-populations. We binned the true and predicted IWI scores into 4 socioeconomic classes[7] separately to derive a confusion matrix. We found that our XGB (CNN) model achieves on average a recall of 83% (73%) in terms of class prediction on all clusters in the test set, see Table 5. Moreover, while our XGB model tends to predict very accurately the poorer in rural areas, and the richer in urban areas, our CNN model predicts best the poor. Note that the 0% recall values in Table 5 are due to the very small number of clusters in each class: 1 urban-poor and 4 rural-upper-middle clusters.

Train/Test time configuration. Using the re-arranged GT locations (and the augmented data for the CNN model), we found that the recency of GT plays an important role in inference. We see in Table 6 that training and testing on 2019 (the most recent data) outperformed the default combined models that merge the surveys from 2016 and 2019 both for training and testing. Further research is needed to mitigate the recency issue not only when multiple years of survey data are combined but also when the most recent survey is not close-to-date.

5 CONCLUSION

In this preliminary study, we have touched upon the opportunities and challenges of inferring high-resolution poverty maps with multimodal data. We addressed some of these issues by proposing regression models that predict IWI scores from seven data sources, and adjust the train/test configurations to correct for noisy ground-truth locations and time mismatch between wealth and features. Our contributions are four-fold: (1) Our proposed re-arranging strategy outperforms the model with noisy locations. (2) We demonstrated the importance of survey recency when features are not from the same survey-year. (3) We found that population and mobility features are the strongest predictors of wealth, providing alone comparable performances when all features are present. (4) Overall, our models achieve high recall at a socioeconomic level. We found that satellite images are the best for predicting the poor, while metadata-features are the best for predicting the rural-poor and the urban-middle classes. As next steps we aim to explore the spatial and temporal transferability of multi-source models with missing data layers at the target population and predict the variability of IWI scores within clusters. Nevertheless, even these preliminary results shed light on the importance of producing accountable models that guarantee accurate predictions to all sub-populations.

[7] [0-25) poor, [25-50) lower middle, [50,75) upper middle, and [75,100] rich. Note that there are no rich places since the max IWI score in the GT is 72.7, see Section 2.

REFERENCES

Jacob Levy Abitbol and Marton Karsai. Interpretable socioeconomic status inference from aerial imagery through urban patterns. *Nature Machine Intelligence*, 2(11):684–692, 2020.

Jacob Levy Abitbol, Márton Karsai, and Eric Fleury. Location, occupation, and semantics based socioeconomic status inference on twitter. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 1192–1199. IEEE, 2018.

JL Abitbol and AJ Morales. Socioeconomic patterns of twitter user activity. *Entropy (Basel, Switzerland)*, 23(6), 2021.

Lishan Adam and Michael Minges. Icts, ldcs and the sdgs: Achieving universal and affordable internet in the least developed countries, 2018.

Areppim AG. Mobile phone market forecast. https://stats.areppim.com/stats/stats_mobilex2017.htm, 2017. Accessed: 2022-02-02.

Nikolaos Aletras and Benjamin Paul Chamberlain. Predicting twitter user socioeconomic attributes with network and language information. In *Proceedings of the 29th on Hypertext and Social Media*, pp. 20–24. 2018.

Kumar Ayush, Burak Uzkent, Kumar Tanmay3 Marshall Burke2 David Lobell, and Stefano Ermon. Efficient poverty mapping from high resolution remote sensing images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 12–20, 2021.

Kumar AyushBurak Uzkent, Burak Uzkent, Marshall Burke, David Lobell, and Stefano Ermon. Generating interpretable poverty maps using object detection in satellite images. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20, 2021. ISBN 9780999241165.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

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), 2022. ISSN 0027-8424. doi: 10.1073/pnas.2113658119. URL https://www.pnas.org/content/119/3/e2113658119.

Eduardo Cruz, Carmen Vaca, and Monica Villavicencio. Estimating urban socioeconomic inequalities through airtime top-up transactions data. In *2021 IEEE International Conference on Big Data (Big Data)*, pp. 4265–4272. IEEE, 2021.

Yuxiao Dong, Yang Yang, Jie Tang, Yang Yang, and Nitesh V Chawla. Inferring user demographics and social strategies in mobile social networks. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 15–24, 2014.

Nathan Eagle, Michael Macy, and Rob Claxton. Network diversity and economic development. *Science*, 328(5981):1029–1031, 2010.

Lisette Espín-Noboa. Povertymaps. https://github.com/lisette-espin/PovertyMaps, 2022. Accessed: 2022-02-02.

Jian Gao, Yi-Cheng Zhang, and Tao Zhou. Computational socioeconomics. *Physics Reports*, 817: 1–104, 2019.

Tilottama Ghosh, Rebecca L Powell, Christopher D Elvidge, Kimberly E Baugh, Paul C Sutton, and Sharolyn Anderson. Shedding light on the global distribution of economic activity. *The Open Geography Journal*, 3(1), 2010.

Google. The maps static api. https://developers.google.com/maps/documentation/maps-static/overview, 2022. Accessed: 2022-02-02.

Neal Jean, Marshall Burke, Michael Xie, W Matthew Davis, David B Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016.

Kamwoo Lee and Jeanine Braithwaite. High-resolution poverty maps in sub-saharan africa. *arXiv preprint arXiv:2009.00544*, 2020.

Yannick Leo, Eric Fleury, J Ignacio Alvarez-Hamelin, Carlos Sarraute, and Márton Karsai. Socioeconomic correlations and stratification in social-communication networks. *Journal of The Royal Society Interface*, 13(125):20160598, 2016.

Yannick Leo, Márton Karsai, Carlos Sarraute, and Eric Fleury. Correlations and dynamics of consumption patterns in social-economic networks. *Social Network Analysis and Mining*, 8(1):1–16, 2018.

Alejandro Llorente, Manuel Garcia-Herranz, Manuel Cebrian, and Esteban Moro. Social media fingerprints of unemployment. *PloS one*, 10(5):e0128692, 2015.

Meta. Data for good. `https://dataforgood.facebook.com/`, 2022. Accessed: 2022-02-02.

Meta. Reach estimate. `https://developers.facebook.com/docs/marketing-api/reference/reach-estimate/`, 2022. Accessed: 2022-02-02.

Franz-Benjamin Mocnik, Amin Mobasheri, and Alexander Zipf. Open Source Data Mining Infrastructure for Exploring and Analysing OpenStreetMap. *Open Geospatial Data, Software and Standards*, 7(3), 2018. doi: 10.1186/s40965-018-0047-6.

OpenStreetMap. Populated places. `https://wiki.openstreetmap.org/wiki/Key:place`, 2021. Accessed: 2022-02-02.

OpenStreetMap. Explorer. `https://www.openstreetmap.org/`, 2022. Accessed: 2022-02-02.

Maxim Pinkovskiy and Xavier Sala-i Martin. Lights, camera… income! illuminating the national accounts-household surveys debate. *The Quarterly Journal of Economics*, 131(2):579–631, 2016.

Jeroen Smits and Roel Steendijk. The international wealth index (iwi). *Social indicators research*, 122(1):65–85, 2015. Accessed: 2022-02-02, Play with IWI: `https://globaldatalab.org/iwi/`.

Statistics Sierra Leone and ICF. Sierra leone demographic and health survey 2019 [dataset] slhr7afl.dta, slge7afl.shp. `https://dhsprogram.com/methodology/survey/survey-display-545.cfm`, 2020. Accessed: 2021-04-16.

Catholic Relief Services Statistics Sierra Leone, University of Sierra Leone and ICF. Sierra leone malaria indicator survey 2016. [dataset] slhr72fl.dta, slge71fl.shp. `https://dhsprogram.com/methodology/survey/survey-display-515.cfm`, 2016. Accessed: 2021-04-16.

The DHS Program. The demographic and health surveys (dhs) program. `https://dhsprogram.com/`, 2022a. Accessed: 2022-02-02.

The DHS Program. Gps data collection. `https://dhsprogram.com/methodology/gps-data-collection.cfm`, 2022b. Accessed: 2022-02-02.

United Nations. Sustainable development goals - goal 1: End poverty in all its forms everywhere. `https://www.un.org/sustainabledevelopment/poverty/`, 2020. Accessed: 2022-02-02.

Unwired Labs. Opencellid - the world's largest open database of cell towers. `https://www.opencellid.org`, 2021. Accessed: 2022-02-02.

Ben Wisner, Piers Blaikie, Terry Cannon, and Ian Davis. *At risk: natural hazards, people's vulnerability and disasters*. Routledge, 2014.

World Bank. Decline of global extreme poverty continues but has slowed: World bank. https://www.worldbank.org/en/news/press-release/2018/09/19/decline-of-global-extreme-poverty-continues-but-has-slowed-world-bank, 2018. Accessed: 2022-02-02.