Mapping Deprived Areas in Low- and Middle-Income Countries (LMIC)

Dr. Ryan Engstrom
Department of Geography
George Washington University
Washington DC USA
rengstro@gwu.edu

Maxwell Owusu
Department of Geography
George Washington University
Washington DC USA
mowusu@gwu.edu

Dr. Amir Jafari
Data Science Program
George Washington University
Washington DC USA
ajafari@gwu.edu

Daqian Dang
Data Science Program
George Washington University
Washington DC USA
daqian@gwu.edu

Akhil Bharadwaj
Data Science Program
George Washington University
Washington DC USA
abharadwaj10@gwu.edu

Dr. Michael Mann
Department of Geography
George Washington University
Washington DC USA
mmann1123@gmail.com

Abstract— In low- and middle-income countries (LMICs), understanding the spatial distribution of deprived areas is crucial for effective policymaking, resource allocation, and targeted interventions aimed at reducing poverty and improving socio-economic conditions. However, traditional methods of identifying deprived areas often rely on outdated or limited data sources, leading to incomplete or inaccurate assessments. To address this challenge, our project proposes developing a comprehensive mapping framework for identifying deprived areas in LMICs using geospatial analysis and machine learning techniques. Leveraging satellite imagery, census data, and other geospatial datasets, our approach aims to create high-resolution maps that capture key deprivation indicators, such as access to basic services, infrastructure, and socio-economic status. In this study, we demonstrate to map deprived areas in two African cities-Lagos (Nigeria) and Nairobi (Kenya). Contextual features and reference files were extracted at 10 m spatial resolution and aggregated into a 100m grid. By integrating advanced spatial analytics and machine learning algorithms, we seek to enhance the accuracy and granularity of deprivation mapping, enabling policymakers and stakeholders to better target resources and interventions to areas in greatest need. The resulting maps will provide valuable insights into the spatial distribution of deprivation, facilitate evidence-based decision-making, and support efforts to achieve sustainable development goals in LMICs.

Keywords-- contextual, covariate, deprived, slums, Lagos, Nairobi

I. INTRODUCTION

The prevalence and growth of slums within low- and middle-income countries (LMICs) have emerged as a considerable challenge, particularly in the COVID-19 pandemic and accompanying economic crises. Slums, often characterized by inadequate housing, overcrowding, and lack of access to clean environments, represent some of the most visible manifestations of urban poverty and inequality. The pandemic has exacerbated those conditions highlighting the vulnerabilities of slum dwellers to health and economic shocks. The economic downturn triggered by the pandemic has further strained the limited resources available for communities, pushing more individuals into conditions of poverty and precarious living situations. This situation presses urgent attention for focused interventions and support to address the challenges faced by slum residents, not only to combat the immediate impacts of the pandemic but also to improve their long-term resilience and well-being.

The accurate mapping of deprived areas, including slums, is crucial for monitoring progress towards the Sustainable Development Goals (SDGs), which aim to make cities and human settlements inclusive, safe, resilient, and sustainable. Detailed mapping provides essential data that can inform policymaking and resource allocation, ensuring that interventions are targeted effectively to address the needs of the most vulnerable populations. It also facilitates the monitoring of changes over time, enabling stakeholders to assess the impact of interventions and adjust strategies accordingly as necessary. Furthermore, accurate mapping supports the identification of gaps in service provision and infrastructure, guiding efforts to improve access to essential services such as clean water, healthcare, and education. Therefore, the task of mapping deprived areas is not just a technical challenge but a critical step towards achieving equity and inclusion in urban development.

II. LITERATURE REVIEW

The integration of big data and geospatial analysis is revolutionizing urban planning and environmental research, presenting a myriad of opportunities alongside formidable challenges. A recent literature review indicates common difficulties such as managing massive and noisy datasets, integrating varied data sources, addressing privacy issues, and requiring advanced technical know-how and cross-disciplinary efforts. Lee and Kang (2015) [1] stress the need for advanced algorithms to process and store the ever-growing and diverse geospatial big data. Robinson et al. (2017) [2] highlight the challenges of big data in cartography, focusing on synthesizing large data volumes into coherent and comprehensible visual formats. Lehner et al. (2020) [3] discuss the hurdles in distilling useful insights from sprawling, unstructured datasets, and the integration challenges within urban environments. Koldasbayeva et al. (2023) [4] explore geospatial modelling in environmental research, tackling issues like imbalanced datasets and spatial autocorrelation, as well as accurately quantifying uncertainty in predictions. Runfola et al. (2024) [5] demonstrate the application of deep learning to estimate socioeconomic

factors from satellite imagery, addressing the variance in geographic scope. Collectively, these works underscore the necessity for sophisticated data management, analytical approaches, and interdisciplinary collaboration to fully exploit geospatial big data to improve urban living conditions and environmental studies.

In the project under discussion, a multifaceted approach is employed, starting with an in-depth geospatial analysis of Lagos and Kano, where factors such as urban infrastructure (e.g., health, schools, and employment opportunities) [6] population density, and other significant geographic elements are explored to understand their impact on urban deprivation. After this analysis, the project will integrate these geospatial insights with sophisticated machine learning and deep learning algorithms. Both machine learning and deep learning mythologies are crucial in this project, as they facilitate the processing of extensive datasets and enable the identification of intricate patterns in urban environments. These approaches significantly enhance the precision and depth of deprivation classification in cities like Lagos and Kano, leveraging the strengths of each method to achieve a more comprehensive analysis. There are some challenges like having imbalanced data with fewer data points for deprived regions [7] which can affect our model's performance. Therefore, data augmentation methods like image flipping, changing brightness and contrast, and increasing synthetical training samples for deprived areas. In user-driven earth observation-based slum mapping [8] Owusu mentions the need for two main elements which are a key to successful slum-based mapping involving a.) contextualizing slums or understanding of local context and user requirements. T. Stark [9] talks about the challenges with imbalanced datasets while using deep learning models and handles them using augmentation techniques like rotation and affine transformations and implements transfer learning for slum mapping. Another challenge is that slum labels change over time in a particular area subject to social, political, and economic changes so Fisher [10] introduces a concept called uncertainty and quantifies it to help us in understanding and detecting slum area changes over a given period.

III. METHODOLOGY

The methodology applied in this study combines data from covariate, contextual, and Sentinel-2-pixel features to analyze and map deprived urban areas in Lagos, Nigeria. The process involves extensive data manipulation, feature extraction, and the integration of various data forms to assess socio-economic conditions and environmental factors.

A. Covariate Features

Our analysis harnesses a dataset from a multi-band raster file encompassing 53 covariate features, aiming to enhance predictive model performance for a nuanced understanding of urban deprivation. We extract data from a GeoPackage file that subdivides Lagos into a 100-meter grid, transforming grid geometry into centroid points. This translates into a CSV output where each row pinpoints a grid cell labelled for deprivation, capturing the complex tapestry of Lagos's landscape.

B. Contextual Features

Contextual features are pivotal for depicting the spatial dynamics of Lagos, offering insights into its structural nuances through variables like building density and climate risk factors. Resampling efforts adjust the resolution of contextual TIFF files to ensure alignment with our spatial framework, enabling a more precise geographic analysis. From these adjusted files, we extract a rich array of spatial data, culminating in a comprehensive dataset that reflects the city's spatial characteristics, enriched by an array of sophisticated features like Fourier Transforms and Histogram of Oriented Gradients.

C. Sentine-2 Satellite Pixel Features

Sentinel-2 satellite imagery provides a rich canvas for our analysis, thanks to its high resolution and versatility. We align this imagery with our spatial framework through resampling, facilitating a detailed examination of Lagos's zones. This imagery allows us to investigate the city's physical attributes deeply, organizing the data by geographic location for an in-depth exploration of environmental and urban dynamics.

D. Data Combination Process

The integration objective is to weave together data from contextual features, covariate bands, and pixel values into a singular dataset that encapsulates the multi-dimensional aspects of Lagos. Starting with meticulously prepared datasets, we align and synthesize this information based on geographic identifiers. The resulting consolidated dataset forms a comprehensive basis for modelling and analysis, providing a holistic representation of Lagos's urban and environmental fabric.

IV. RESULTS

The dataset used for training for Lagos contained 305, 381 rows with 204 columns that had 4 features for pixel values obtained from Sentinel-2 images, 53 covariate features and 144 contextual features, longitude, and latitude. There slum labels contained values 0, 1, 2 and 3 (see Fig.1). The label values 1 and 2 were combined as a single slum value 1 that indicates deprived area. While 0 as indicated in Fig4. was considered as a non-deprived area. Label value 3 was added to test set for modelling as it was unsure. For training purposes, a Python package called PyCaret was used. A default train test split of 70% for training and 30% for testing was used.

Figures and Data Tables

Value	Classification
0	This is no slum. This is the baseline we created above when we created a new column called Slum.
1	This is a slum (95% coverage). Verified as far possible on Google Street View and by Adenike. This label made sure that there would be no interference from roads or anything green for example if it was used for training data.
2	This is 50% slum. We included this as we had previously discounted any areas, we thought would affect the model. It was decided that slums do contain areas of water, green, 'built' roads, and formal-looking buildings (rather than just irregular layouts). We knew the locations of the slums but weren't able to say that this would be the best data if used for training.
3	This was used by Alex in areas she was unsure of. We had over 300,000 features so 3 was a way of putting a 'post-it note' on an area for Alex to ask for verification.

Fig. 1 Description of target variable Slum

Model	Accuracy	AUC	Recall	Precision	F1	Kappa	MCC
Logistic Regression	0.9978	0.9477	0.085	0.3953	0.1244	0.124	0.1975
Random Forest Classifier	0.9986	0.9967	0.44	0.9263	0.5772	0.5766	0.6138
K Neighbors Classifier	0.9978	0.5065	0	0	0	0	0
XGBoost	0.9988	0.9994	0.595	0.8322	0.6706	0.67	0.6807
Multi-Layer Perceptron	0.9978	0.6787	0	0	0.1408	0.1392	0.1413
Decision Tree	0.9982	0.7946	0.52	0.5683	0.5915	0.5906	0.5906
Ridge Classifier	0.9978	0.5	0	0	0	0	0
Quadratic Discriminant Analysis	0.9732	0.9873	0.995	0.0753	0.139	0.1355	0.2683
Ada Boost Classifier	0.9985	0.9942	0.53	0.7413	0.6205	0.6198	0.6234
Gradient Boosting Classifier	0.9978	0.742	0.265	0.6625	0.1626	0.1619	0.2077
Extra Trees Classifier	0.9986	0.9993	0.405	0.9205	0.5782	0.5776	0.6194
Light Gradient Boosting Machine	0.9956	0.9174	0.435	0.1608	0.3228	0.3208	0.3405
Dummy Classifier	0.9978	0.5	0	0	0	0	0
Support Vector Machine	0.9837	0.4954	0.01	0.0161	0.0013	0	0
Naive Bayes	0.0686	0.8725	0.995	0.0023	0.0047	0.0003	0.0125

Table 1: Classical modelling results on test set using imbalanced data for Lagos.

Model	Accuracy	AUC	Recall	Precision	F1	Kappa	MCC
Logistic Regression	0.9141	0.9602	0.885	0.0231	0.0449	0.0408	0.1384
Random Forest Classifier	0.9986	0.9964	0.605	0.5789	0.6471	0.6463	0.6479
K Neighbors Classifier	0.9417	0.5604	0.17	0.0062	0.0111	0.0069	0.0188
XGBoost	0.9989	0.9986	0.73	0.7807	0.7404	0.7398	0.7401
Multi-Layer Perceptron	0.9692	0.9839	0.855	0.0773	0.1171	0.1134	0.2373
Decision Tree	0.9976	0.8067	0.54	0.383	0.5302	0.529	0.5341
Ridge Classifier	0.9688	0.9769	0.98	0.0594	0.1213	0.1177	0.2482
Quadratic Discriminant Analysis	0.9849	0.9878	0.98	0.1134	0.2193	0.2163	0.3435
Ada Boost Classifier	0.9943	0.9951	0.81	0.2278	0.363	0.3609	0.4208
Gradient Boosting Classifier	0.9945	0.9983	0.94	0.2517	0.4208	0.4189	0.4974
Extra Trees Classifier	0.9988	0.9991	0.75	0.6977	0.7172	0.7166	0.7166
Light Gradient Boosting Machine	0.9987	0.9991	0.745	0.6835	0.7041	0.7034	0.7036
Dummy Classifier	0.9978	0.5	0	0	0	0	0
Support Vector Machine	0.8667	0.801	0.785	0.0128	0.0235	0.0193	0.0824
Naive Bayes	0.0759	0.577	0.995	0.0023	0.0046	0.0003	0.0096

Table 2: Classical modelling results on test set using balanced data for Lagos.
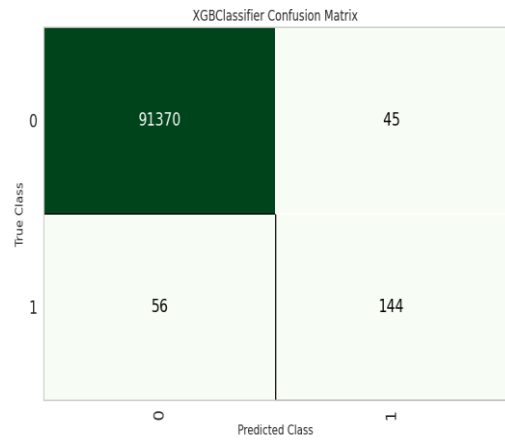

Fig.2. XGBoost classifier confusion matrix. 0 represents non-deprived and 1 is for deprived
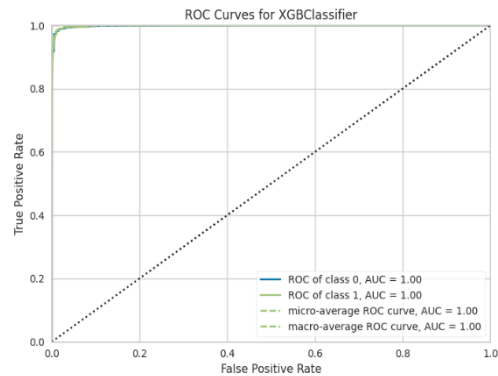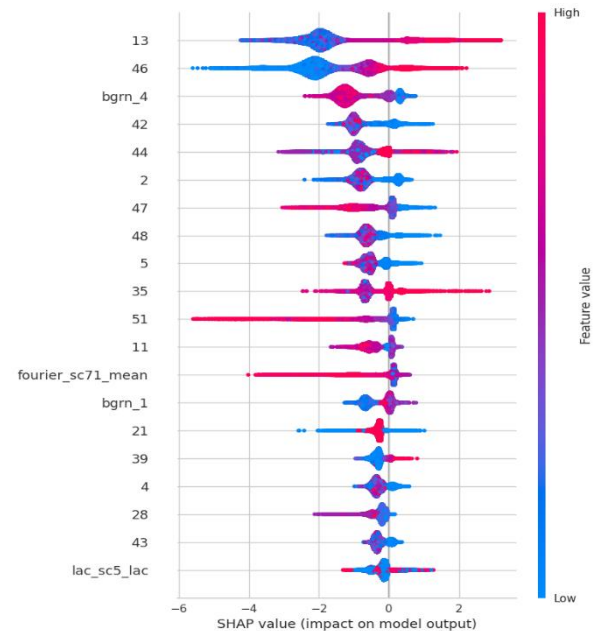

Fig.3. ROC curve for XGBoost clas

Fig.4. SHAP values showing impact of features on model output

Covariate Feature	Description
13	ph_dist_cultivated_2015
46	uu_bld_den_2020
42	sh_dist_mnr_pofw_2019
44	sh_ethno_den_2020
2	fs_dist_school_2020
47	uu_impr_housing_2015
48	fs_dist_hf_2019
5	in_dist_waterway_2016
35	ses_m_lit_2018
51	ph_dist_riv_network_2007
11	ph_dist_art_surface_2015
21	ph_grd_water_2000
39	ses_preg_2017
4	in_dist_rd_intersect_2016
28	ph_slope_2000
43	sh_dist_pofw_2019

Table 3: Covariate feature description based on Fig. 4
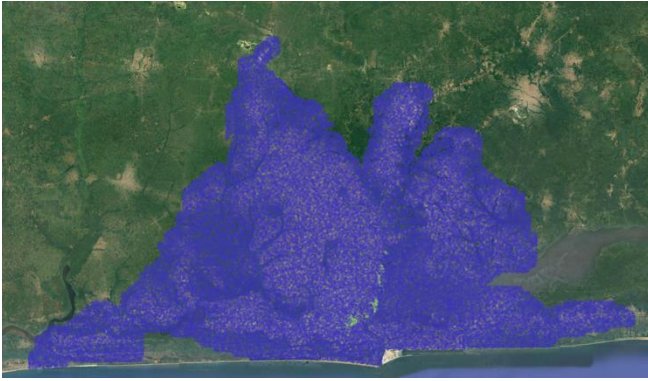
Fig. 5: A test set result for Lagos after mapping the points on QGIS software

In Figure 5, the points in blue represent non-deprived areas and the green points represent deprived areas. XGBoost classifier was used which obtained an F1 score of 0.75 to generate predictions on the test set for Lagos. These predictions were then exported to a CSV file and were imported into QGIS software. The coordinates (longitudes and latitudes) were then used to map the points on QGIS software by adding a layer and using the EPSG:4326 (WGS 84) coordinate reference system. Google Earth satellite image was added as a layer and the points were classified based on the prediction labels obtained from the modelling results into deprived and non-deprived which were mapped on QGIS software.

V. Discussion

Our results have validated the robustness of our mapping framework, employing machine learning models, notably XGBoost, which demonstrated commendable precision in pinpointing deprived urban sectors. The intricate analysis of the models' performance, delineated by accuracy metrics and the elucidation of influential features through SHAP values, reflects their potential utility. This study navigated through a labyrinth of data processing complexities, stemming primarily from the heterogeneity of geospatial data formats, which was adeptly handled by the integration capabilities of advanced libraries, enhancing our pipeline's capacity for feature extraction and data synthesis.

However, the project faced significant hurdles in harmonizing the data due to variances in spatial resolution. The meticulous process of downscaling GeoPackage files to a finer 10-meter resolution was undertaken to bolster our analytical granularity, allowing for the application of sophisticated machine learning and deep learning methodologies. This higher-resolution data capture enabled the distillation of more detailed spatial features, augmenting the predictive acumen of our models and their transferability across different urban landscapes. The classical modelling approach achieved a promising baseline performance, signalling the possibility for even greater enhancements with the inclusion of deep learning constructs such as Convolutional Neural Networks (CNNs).

The scope of our work, while demonstrating significant potential, invites further exploration into the resampling of data at finer resolutions and its impact on model fidelity. The ambition is to sculpt a model with the agility to generalize its predictive capabilities, transcending the unique urban contours of Lagos to apply its analytical prowess to a broader canvas of cities worldwide. Such adaptability is imperative for the global extrapolation of our models, aiming to empower policymakers and stakeholders with a scalable tool for urban analysis and planning.

VI. Conclusion

The project showcases significant advancement in understanding and mitigating urban poverty, particularly through geospatial slum mapping in developing nations. Despite data diversity and resolution challenges, the project established a refined data pipeline with tools like GeoWombat and resampling strategies. With an F1 score of 0.75 from a classical machine learning model, there is scope for enhancement using advanced deep learning and data refinement. The project's models are scalable and can be extended from Lagos and Nairobi to other cities globally, aligning with the Sustainable Development Goals for inclusive, safe, resilient, and sustainable cities. The progress signals a positive future for urban development and poverty reduction, emphasizing the crucial role of precise mapping of deprived areas in social justice and equitable urban planning.

VII. Acknowledgement

VIII. REFERENCES

[1] Lee, J.-G., & Kang, M. (2015). Geospatial Big Data: Challenges and Opportunities. *Big Data Research*, *2*(2), 74–81. https://doi.org/10.1016/j.bdr.2015.01.003

[2] Robinson, A. C., Demšar, U., Moore, A. B., Buckley, A., Jiang, B., Field, K., Kraak, M.-J., Camboim, S. P., & Sluter, C. R. (2017). Geospatial big data and cartography: research challenges and opportunities for making maps that matter. *International Journal of Cartography*, *3*(sup1), 32–60. https://doi.org/10.1080/23729333.2016.1278151

[3] Kovacs-Györi, A., Ristea, A., Havas, C., Mehaffy, M., Hochmair, H. H., Resch, B., Juhasz, L., Lehner, A., Ramasubramanian, L., & Blaschke, T. (2020). Opportunities and Challenges of Geospatial Analysis for Promoting Urban Livability in the Era of Big Data and Machine Learning. *ISPRS International Journal of Geo-Information*, *9*(752).

[4] Koldasbayeva, D., Tregubova, P., Gasanov, M., Zaytsev, A., Petrovskaia, A., & Burnaev, E. (2023). Challenges in data-based geospatial modeling for environmental research and practice. *arXiv.Org*. https://doi.org/10.48550/arxiv.2311.11057

[5] Runfola, D., Stefanidis, A., Lv, Z., O'Brien, J., & Baier, H. (2024). A multi-glimpse deep learning architecture to estimate socioeconomic census metrics in the context of extreme scope variance. *International Journal of Geographical Information Science*, 1–25. https://doi.org/10.1080/13658816.2024.2305636

[6] Engstrom, R., Owusu, M., Nair, A., Jafari, A., Thomson, D., & Kuffer, M. (2023). Evaluating the ability to use contextual features to map deprived areas 'slums' in multiple cities. *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*. https://doi.org/10.1109/igarss52108.2023.10282954

[7] Owusu, M., Engstrom, R., Thomson, D., Kuffer, M., & Mann, M. L. (2023, November 8). *Mapping deprived urban areas using open geospatial data and machine learning in Africa*. MDPI. https://www.mdpi.com/2413-8851/7/4/116/htm

[8] Owusu, M., Kuffer, M., Belgiu, M., Grippa, T., Lennert, M., Georganos, S., & Vanhuysse, S. (2021). Towards user-driven earth observation-based slum mapping. *Computers, Environment and Urban Systems*, *89*, 101681. https://doi.org/10.1016/j.compenvurbsys.2021.101681

[9] T. Stark, M. Wurm, H. Taubenböck and X. X. Zhu, "Slum Mapping in Imbalanced Remote Sensing Datasets Using Transfer Learned Deep Features," *2019 Joint Urban Remote Sensing Event (JURSE)*, Vannes, France, 2019, pp. 1-4, doi: 10.1109/JURSE.2019.8808965.

[10] Fisher, T., Gibson, H. W., Liu, Y., Abdar, M., Posa, M., Salimi-Khorshidi, G., Hassaïne, A., Cai, Y., Rahimi, K., & Mamouei, M. (2022). Uncertainty-Aware interpretable deep learning for slum mapping and monitoring. *Remote Sensing*, *14*(13), 3072. https://doi.org/10.3390/rs14133072