Auditing Geospatial Datasets for Biases: Using Global Building Datasets for Disaster Risk Management

Caroline M. Gevaert, Thomas Buunk, and Marc J.C. van den Homberg

***Abstract*—The presence of biases has been demonstrated in a wide range of machine learning applications; however, it is not yet widespread in the case of geospatial datasets. This study illustrates the importance of auditing geospatial datasets for biases, with a particular focus on disaster risk management applications, as a lack of local data may direct humanitarian actors to utilize global building datasets to estimate damage and the distribution of aid efforts. It is important to ensure that there are no biases against the representation of vulnerable populations and that they are not missed in the distribution of aid. This manuscript audits four global building datasets [Google Open Buildings, Microsoft Bing Maps Building Footprints, Overture Maps Foundation (OMF), and OpenStreetMap (OSM)] for biases regarding the relative wealth index (RWI), population density, urban/rural proportions, and building size in Tanzania and the Philippines. The dataset accuracies for these two countries are lower than expected. Google Open Buildings (with a confidence above 0.7) and OSM demonstrated the best combinations of false negative and false discovery, though Google Open Buildings was more consistent across tiles. The equality of opportunity was lowest for the urban/rural proportions, whereas the OSM and OMF displayed particularly low equality of opportunity for population density and RWI in Tanzania. These results demonstrate that biases exist in these geospatial datasets. The types of biases are not consistent across the datasets and the two study areas, which emphasizes the importance of auditing these datasets for biases in new applications and study areas.**

***Index Terms*—Bias, building detection, equity, ethics, humanitarian aid, machine learning.**

I. Introduction

OVER the last few decades, the number of disasters and their impacts have increased. Disasters have different impacts, which have been observed along wider axes of societal marginalization, such as income and social status, age, race, gender, and disability [1]. Therefore, disaster risk management interventions aim to realize procedural, recognition, and distributional equity. For example, humanitarian actors aim to support those who are the most vulnerable and to do this impartially, without any discrimination or bias [2]. Accounting for equity places stringent requirements on geospatial data and methods.

Geospatial data and information are important for supporting the decision-making processes in disaster risk management. In the preparedness phase, geospatial data are used to assess the different dimensions of risk, that is, hazard, exposure, and vulnerability. In addition, geospatial data were used to train and run predictive models for anticipatory action. During the response phase, it is essential to assess the impact of disasters. Geospatial data must have sufficiently high spatial and temporal resolution and coverage for these applications.

All disaster risk management phases require exposure data on infrastructure. Information regarding buildings is crucial for detecting areas that are susceptible to disasters or those affected by a disaster. National Statistics Offices and Land Registries hold data on population and building density. However, these data are often collected at long intermittent intervals and disclosed only at an aggregated level. The Digital Divide is one of the factors that makes it especially difficult for developing countries to create and maintain good databases of exposure data [3], [4].

Simultaneously, significant advances in artificial intelligence (AI) and the growing availability of big data, particularly Earth Observation data, have led to global initiatives with the potential to overcome the problem of limited data in such countries. In particular, big tech companies have released open building delineation datasets, such as Google's Open Buildings [5], Microsoft's Building Footprints [6], [7], and the Overture Maps Foundation (OMF) [8], which cover a significant amount of the globe. These building datasets were created using specific deep-learning AI algorithms, whereby biases were introduced at various stages [9]. There are representation biases, which refer to biases induced by the data generation process or data collection. Biases arise when the sample does not accurately represent the population. Algorithmic biases can also occur when an algorithm, for example, performs better on one type of built-up area than on another. These biases can have a direct influence on the equity of interventions that use AI.

Manuscript received 2 January 2024; revised 26 April 2024 and 18 June 2024; accepted 24 June 2024. Date of publication 3 July 2024; date of current version 24 July 2024. The work of Caroline M. Gevaert was supported by the project "Bridging the Gap Between Artificial Intelligence and Society: Developing Responsible and Viable Solutions for Geospatial Data" (with project number 18091) of the research program Talent Programme Veni 2020 by the Dutch Research Council (NWO). The work of Marc J.C. van den Homberg was supported by the project "Compound Flood Inundation: DE_Compound Operational Forecasting and Assessment of Climate Risks," of the ECWMF Destination Earth Programme. *(Corresponding author: Caroline M. Gevaert.)*

Caroline M. Gevaert is with the Faculty ITC, University of Twente, 7522 NB Enschede, The Netherlands (e-mail: c.m.gevaert@utwente.nl).

Thomas Buunk is with the 510 An Initiative of the Netherlands Red Cross, 2593 HT The Hague, The Netherlands (e-mail: Johnthomasbuunk@outlook.com).

Marc J.C. van den Homberg is with the Faculty ITC, University of Twente, 7522 NB Enschede, The Netherlands, and also with 510 An Initiative of the Netherlands Red Cross, 2593 HT The Hague, The Netherlands (e-mail: mvandenhomberg@redcross.nl).

Digital Object Identifier 10.1109/JSTARS.2024.3422503

Unintended biases regarding gender and race have been observed in AI algorithms [10] for image recognition [11], [12], credit scoring [13], [14], and criminal recidivism [15]. It is logical that such biases are also present in algorithms used for disaster response or humanitarian aid. Building footprint datasets are used for purposes such as population estimation [16] and disaster risk management [17], [18], [19], poverty estimation [20], and the identification of vulnerable population groups [21]. However, there is a lack of research on how biases manifest themselves in building datasets and the potential implications for these applications. For example, what if AI algorithms that delineate buildings perform worse in informal areas of a city than in formal areas? What if the results of this algorithm were used to distribute aid after a flooding event? We will actually be missing people who need it the most.

Auditing global building footprint datasets for biases has not been done before and is often limited to assessing completeness and comparing datasets. For example, one study assessed the spatial variation of OpenStreetMap (OSM) completeness globally [22]. Another study compared building outlines from Google, Ecopia, Microsoft, and OSM across Africa [23], but due to a lack of reference data, it simply compared the datasets and identified large discrepancies in the estimated building counts. It did not quantify the dataset correctness or completeness, nor did it assess potential biases. Kuffer et al. [24] compared the completeness of gridded population datasets between high-, low-, and middle-income countries and found that these population datasets have significantly more errors in low- and middle-income countries than in high-income countries. These studies do not relate global building footprint datasets to social biases that may have implications in their usage for development or disaster risk management in particular.

This study investigated biases in global datasets of building outlines through two case studies: one in Tanzania and the other in the Philippines. This demonstrates how to audit geospatial datasets for fairness and biases and links literature from ethical AI with applications in the geospatial domain. It is very relevant for Global South applications, where the lack of local data often causes GIS experts to turn to global datasets, which may not always have been validated locally. Section II conceptualizes the notion of bias by drawing on global guidelines and academic literature. Section III explains the case studies, datasets, and audit methodology. The results are presented in Section IV and discussed in Section V. Finally, Section VI concludes this article.

II. Conceptualization and Quantification of Bias

UNESCO released its “First Draft of the Recommendation on The Ethics of Artificial Intelligence” in September 2020 [25]. It does not specifically define bias, though it states: “AI actors should make all efforts to minimize and avoid reinforcing or perpetuating inappropriate sociotechnical biases based on identity prejudice throughout the life cycle of the AI system to ensure fairness of such systems. There should be a possibility of having a remedy against unfair algorithmic determination and discrimination”. The Recommendations further emphasize the need for a globally accepted Ethical Impact Assessment as well as the development of mechanisms to audit AI technologies regarding their compliance with the stated principles and values.

The European Commission goes beyond recommendations and is drafting a binding Regulation for an Artificial Intelligence Act [26] with the aim of ensuring “the development, use, and uptake of AI in the internal market that at the same time meets a high level of protection of public interests, such as health and safety and the protection of fundamental rights.” The Act intends to regulate AI algorithms to ensure they do not violate fundamental rights. Datasets used for the development of AI algorithms must be examined for possible biases and any gaps or shortcomings must be addressed.

Assessing biases in machine learning has been extensively researched in the field of Trustworthy AI, also known as Responsible AI or Ethical AI. Many definitions of fairness and methods to audit and mitigate biases in algorithms have been developed in the past few years [27], [28], [29], [30]. A common approach to assessing group biases is to use statistical fairness criteria to compare the results of a classification model for the various subgroups of the population. These subgroups are determined by varying values of a so-called “sensitive attribute.” For example, if the sensitive attribute is economic income, the subgroups represent richer parts of the population to poorer, vulnerable, groups of the population. If the algorithm systematically performs worse for vulnerable groups compared to richer groups, then it is potentially bias. The statistical fairness criteria measure how large these systematic differences are. These statistical measures are generally grouped into three categories: independence, separation, and sufficiency [27].

Let us define a classification task that takes a set of features X and predicts a class label $\hat{Y}$. Y is the actual class of the sample and is the ground truth value. Furthermore, we have A as the group attribute, that is, the sensitive characteristic of the sample. Note that although A can represent any type of group, when auditing for biases, it represents attributes related to groups that are considered to be sensitive or protected and are of particular interest (e.g., gender, ethnicity, vulnerable socioeconomic groups). *Independence*, also referred to as demographic parity or statistical parity, considers that the output classification should be independent of sensitive attributes:

$$\hat{\mathrm{Y}} \perp A. \tag{1}$$

To achieve independence, the error rates of the classifier should be the same for all variables. In practice, this constraint is often relaxed by using a slack variable $\in$. Thus in a binary classification scheme with $Y = \{0, 1\}$ and a sensitive attribute with two subgroups $A = \{a, b\}$, a relaxed interpretation of independence would be [27]

$$\frac{P(\hat{Y} = 1|A = a)}{P(\hat{Y} = 1|A = b)} \geq 1- \in . \tag{2}$$

Separation considers the error rates of the classifiers for various groups. It states that the output classification should be conditionally independent of the sensitive attribute for each subgroup

$$\hat{Y} \perp A \mid Y. \tag{3}$$

It thus considers both false positives (FPs) and false negatives (FNs), and rather than demanding that a model has an equal accuracy rate for all subgroups, it demands that the FP and FN rates are balanced. This is also referred to as *equality of odds* [31]. Balancing the FP rate across groups is known as predictive equality (4) and balancing FN is known as *equality of opportunity* (5)

$$\frac{P(\hat{Y} = 1|A = a, Y = 0)}{P(\hat{Y} = 1|A = b, Y = 0)} \geq 1 - \epsilon \tag{4}$$

$$\frac{P(\hat{Y} = 0|A = a, Y = 1)}{P(\hat{Y} = 0|A = b, Y = 1)} \geq 1 - \epsilon. \tag{5}$$

The final category of statistical measures, sufficiency, considers model precision given the predicted outcome rather than the true value

$$Y \perp A \mid \hat{Y}. \tag{6}$$

Ensuring sufficiency is also known as *predictive parity* [32]. The relationship between separation and sufficiency is similar to that between recall, i.e., $P(\hat{Y} = 1|Y = 1)$, and precision, i.e., $P(Y = 1|\hat{Y} = 1)$ [28]. Predictive parity can be formulated as

$$\frac{P(Y = 1|A = a, \ \hat{Y} = 1)}{P(Y = 1|A = b, \ \hat{Y} = 1)} \geq 1 - \epsilon. \tag{7}$$

Again, the same formulation should be applied to the case of $Y = 0$.

Unfortunately, it is impossible to satisfy all three categories of fairness criteria simultaneously for most cases [27], [28]. For example, if $\hat{Y}$ is not independent of Y, as is often the case in classifiers, then independence and separation cannot be satisfied simultaneously. Furthermore, for binary classification problems, the presence of a single FP ensures that separation and sufficiency are incompatible [28], [33]. Selecting an appropriate fairness metric depends on the application, philosophical understanding of fairness, and perceived harm of different forms of bias. Some tools have been developed to guide users in the selection of the most appropriate metrics [34].

Fairness metrics are often used for binary classification problems with one sensitive attribute and two groups. Indeed, this is how they are presented in this section. When a sensitive attribute has multiple groups, (2), (4), and (5) can be reformulated to represent the group with the minimum probability in the numerator and the group with the maximum probability in the denominator [35], [36], [37].

III. Methodology

The overall workflow consisted of

1) selecting datasets (the datasets to be audited, datasets representing sensitive attributes, and datasets serving as reference data),
2) calculating the accuracy of each dataset, and
3) determining discrepancies in the accuracy metrics across sensitive attributes through selected bias metrics. The workflow was demonstrated through two case studies: Tanzania and the Philippines.

A. Case Studies

1) Tanzania: Tanzania is classified as a lower middle income country. It had an estimated GDP of 1192.8 USD, a population density of 72 persons per square kilometer, and an estimated 37% of the population living in urban areas [38]. The case study in Tanzania is not linked to a specific project but to the context of the widespread use of global building datasets to rapidly quantify exposure by global development agencies. Owing to the lack of up-to-date geospatial data and rapid urbanization, obtaining up-to-date information on the number and distribution of buildings is challenging. Therefore, organizations such as the World Bank have turned to global datasets derived from earth observation to estimate poverty [39] and exposure to hazards [40]. The analysis in Tanzania aimed to identify potential biases for such applications.

2) Philippines: The Philippines is classified as a lower middle income country. It had an estimated GDP of 3498.5 USD, a population density of 382 persons per square kilometer, and an estimated 48% of the population lives in urban areas [38]. The Philippine Red Cross uses a Typhoon Impact-based Forecasting Model as a critical tool in its Early Action Protocol for typhoons in the Philippines. It predicts the potential housing damage of a typhoon before it makes landfall and is used to allocate funding and trigger early actions to reduce the impact [41]. While the initial model predicted housing damage at the municipality level using a vector-based approach, the model was transformed to operate on a 0.1° grid resolution [42]. This shift to a grid-based configuration allows for the integration of globally available datasets, making the model more generalizable and transferable to other countries. The Google Building Footprint dataset played an important role in this transformation. It was used to disaggregate the target variable (percentage of completely damaged houses) and other global features to a finer grid resolution. This was achieved by counting the number of buildings in each grid cell and normalizing it by the total number of buildings in the corresponding municipality [42]. Including the Philippines in the analyses, thereby assessing the appropriateness of using the Google Building Footprint dataset.

B. Datasets

1) Building Datasets to be Audited: Four building datasets were audited for potential biases. The first is OSM [43], a community-driven open-data platform that motivates citizens to digitize buildings in aerial and satellite imagery. Tanzania has been the focus of many humanitarian mapping efforts [44], including the Ramani Huria project, where community members and local university students mapped the city of Dar es Salaam from 2014 to 2019 as part of the Tanzania Urban Resilience Program of the World Bank [45]. The OSM dataset utilized in this study was downloaded from the humanitarian data exchange (HDX) platform [46]. Humanitarian mapping efforts are vital

TABLE I
OVERVIEW OF SENSITIVE ATTRIBUTES UTILIZED FOR THIS STUDY

Variable	Variable name	Year	Source
Relative Wealth index	RWI	2023	Meta's Data for Good
Population density	Pop_dens	2019	Facebook Connectivity Lab, CIESIN and Columbia University
Urban/rural proportion	Rural_scale	2021	Global Urban Rural Catchment Areas by Food and Agriculture Organization (FAO) of the UN
Building size	Bld_size		From reference data

in the Philippines, especially during natural disasters. When Typhoon Mawar hit the Province of Isabela, a dedicated mapping task was established. Similarly, regions affected by floods have frequently become subjects of focused mapping efforts in the country [47].

The second dataset was the Microsoft Bing Maps Building Footprints dataset for Tanzania [6] and the Philippines [7]. In this study, we refer to this dataset as Bing Maps. This dataset contains building footprints generated from Maxar Technologies satellite imagery from 2020 to 2021 using deep learning. Their workflow used a training set of 1.2 million buildings to train an EfficientNet B3. The results of this semantic segmentation are then processed into building outlines. The reported accuracy metrics of the Building Footprint dataset were a precision of 94.5%, a recall of 61.8%, and an Intersection over Union (IoU) of 0.68. These metrics were calculated from an evaluation set of 18 500 buildings in Tanzania and Uganda. From a sample of 6000 buildings across the Philippines, Malaysia, and Indonesia, the precision was reported at 88.6%, recall was 77.5%, and IoU was 65.5%. In addition, FPs, based on a sample of 18 851 buildings in the Philippines, were reported to be 1.8%.

The third dataset is Google Open Buildings [5]. This was another AI-generated building dataset. This model used 1.67 million building polygons spread across Africa, with a concentration around Nigeria and Kenya as training samples. The deep learning model is a residual decoder block inspired by U-Net. The building outlines were then generated using a contouring algorithm. In addition, each building in the Google dataset was assigned a confidence score, reflecting the algorithm's predicted likelihood that a pixel represents a building. This study incorporates an additional dataset, Google Buildings, filtered by a confidence threshold of 0.7, to evaluate its impact on the results. This dataset is referred to as Google_conf throughout the report. Technical details of the training and implementation can be found in [48].

The fourth dataset is from the OMF [8], a collaboration between various mapping, technical, and geospatial companies, including Microsoft and Amazon Web Services. The OMF layer was developed by combining various open-data projects, including OSM, Bing Maps, and ESRI [8]. Data are acquired through the SQL series, as described in the company's GitHub repository [49], with polygons saved in the WKT format. Note that, in the Philippines case study, the OMF layer was identical to the OSM layer and was therefore not included in that part of the analysis.

2) Datasets for Sensitive Attributes: Four sensitive attributes were defined: the relative wealth index (RWI), population density, urban/rural population, and building size (see Table I). They were selected to ensure that they would cover both the Philippines and Tanzania case studies. The RWI dataset predicts the relative standard of living within a country. It was developed by Meta's Data for Good team in collaboration with researchers from the University of California and Berkeley. It uses global household wealth data, collected through face-to-face surveys from villages around the world. Through spatial markers, these villages are linked to data sources, such as satellite imagery, cellular network data, connectivity data from Facebook, and topographic maps. With deep learning and machine learning models, relative wealth is predicted for a 2.4 km^2 cell on the entire populated planet [50]. The dataset can be downloaded per country from HDX and comes in CSV format with latitude, longitude, and RWI values [51].

The population density dataset is a collaborative effort from the Facebook Connectivity Lab, the Center for International Earth Science Information Network (CIESIN), and Columbia University. This dataset was developed by leveraging census data to model population growth at both national and subnational levels. Using satellite imagery, the size and quantity of buildings within 30 × 30 m tiles were estimated. This, in combination with census information, results in population density estimation within these tiles [52]. The dataset was downloaded from the HDX and is available in raster format [53].

The Global Urban Rural Catchment Areas Grid file comprises 30 urban-rural categories based on population distribution, population density, urban center locations, and travel time to urban centers. Category 1 represents a large city with more than 5 million inhabitants, category 30 is the most rural area where travel time to any city is longer than 3 h. The data are available on the website of the Food and Agriculture Organization in raster format [54]. The dataset was at the global level, with a resolution of approximately 900 m.

The sensitive attributes (see Table I) were divided into subgroups to calculate fairness metrics. Here, quantiles were used to automatically select thresholds between different groups. The quantiles of each variable at the national level were calculated and used to obtain four subgroups for each sensitive attribute.

3) Reference Datasets: The reference building outlines for both study areas were generated by manual digitization. The following workflow was used to select reference areas for Tanzania. One hundred points were randomly selected, taking into account the spread over each possible poverty and settlement combination to ensure that each was represented. There are fewer examples of combinations for low settlement size (<10 000 and 10 000–100 000) and poverty levels below 15%, as these were not represented in the data. A square of 250 × 250 m was generated around each point to create the sample tile for further analyses. One hundred tiles were selected to compare the accuracy of the building datasets. These tiles were selected by first identifying the built-up extent of Tanzania using the World Settlement Footprint for 2019 [55]. A regular grid of 5630 points was generated over the built-up areas. Next, the location's poverty and settlement size levels (as defined in Table I) were assigned to each point. Note that not all possible poverty and settlement size classes are equally represented.

For the Philippines, a stratified sampling strategy was implemented for the selection of the study areas. This approach incorporates four key variables at the municipality level: poverty, population density, vulnerable population, and urban/rural proportion. Initially identified as sensitive variables in preliminary research, these variables were adjusted in the current study. Municipalities were segmented into quantiles based on these variables, culminating in the selection of 20 municipalities. Within each of the selected municipalities, a grid composed of tiles measuring 250 × 250 m was established. The subsequent selection of tiles adhered to specific criteria; only those containing a minimum of 50 buildings and featuring at least two of the building datasets under analysis were included. Tiles that did not meet these criteria were excluded. In addition, based on the sensitive variables available at the grid level, stratified sampling was applied again to select tiles within the municipality. Finally, tiles lacking sufficient satellite imagery were removed, resulting in a final dataset of 106 tiles.

For both the Tanzania and Philippines case studies, reference data were generated by manually digitizing the building outlines. This was done in QGIS (version 3.34.1) and using the Google satellite basemap data other areas were digitized using Tasking Manager and ArcGIS Pro v.3.0.2. The building datasets used in this analysis were most likely generated using different image sources. There could be offsets due to differences in the georeferencing of the underlying imagery. Errors may also occur due to the temporal difference between the imagery used to generate the reference data and the building datasets. This study assumed that these temporal differences are negligible. This reference dataset of building outlines was also used to determine the fourth variable used as a possible sensitive attribute. This is because small roof sizes are often perceived as a physical characteristic of slum areas visible in satellite imagery [56].

C. Accuracy Analyses and Auditing for Bias

The analysis was conducted using Python, with inputs including the analyzed and reference-building datasets and the shapefile of tiles containing sensitive variable data. The process involved iterating over tiles and clipping building datasets to the geometry of each tile. Buildings were classified as true positive (TP), FP, or FN based on their intersection with buildings in the reference dataset and the IoU metric. An IoU threshold of 0.5, determined after the initial accuracy assessment was used for classification. The use of 0.5 as the threshold follows the precedence of other researchers [57], [58]. The analysis output included Excel files for each dataset (Bing, Google, and OSM) with building IDs, IoU values, and classification data

$$\mathrm{IoU} = \frac{\left|Y \cap \hat{Y}\right|}{\left|Y \cup \hat{Y}\right|}. \tag{8}$$

To audit for biases, a fairness metric must be selected. In this case, the equality of opportunity (5), also known as the FN rate parity [59] was selected. The use case at hand is to audit building algorithms to understand whether they are systematically missing vulnerable groups of the population. When using these building datasets for humanitarian or development purposes, such as distributing aid or estimating the population located in hazardous areas, we would not want to systematically miss these groups.

At the building level, two fairness metrics, namely, the FN rate (FNR) and false discovery rate (FDR), are utilized for this research, with their respective formulas as follows:

$$\mathrm{FNR} = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TP}} \tag{9}$$

$$\mathrm{FDR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TP}}. \tag{10}$$

The FNR highlights the proportion of positive instances incorrectly classified as negative, which is particularly relevant in situations where missing positive instances carry higher risks or costs than erroneously identifying negative instances as positive. FDR, on the other hand, focuses on the proportion of FPs in all positive predictions. This is a critical metric in scenarios in which false-labeling positives have significant consequences. Note that both low FNRs and FDRs indicate higher accuracy of the model.

Finally, we consider the number of buildings predicted per tile. This may help to lower the sensitivity to false classifications resulting from georeferencing errors. Furthermore, applications such as typhoon impact-based forecasting models use building data aggregated to a grid level. In such cases, the exact building location is not considered; however, the total number of buildings in the grid cell is. To quantify this, we used the normalized building count (NBC), determined as follows:

$$\mathrm{NBC} = \frac{\mathrm{BC}_a - \mathrm{BC}_r}{\mathrm{BC}_r} \tag{11}$$

where BC_a refers to the building count in the dataset being analyzed and BC_r refers to the reference building dataset.

IV. Results

A. Overall Results

Overall, the accuracies of the three building datasets were notably low (see Table II). The FDR varied from 0.22 (OSM) to 0.33 (Bing) in Tanzania, and from 0.23 (OSM) to 0.47 (Google)

TABLE II
KEY ACCURACY STATISTICS OF THE BUILDING DATASETS (TZ = TANZANIA; PH = PHILIPPINES)

Country	Dataset	IoU	TP	FP	FN	FNR	FDR	Building count - national	Building count – study area
Tz	Bing	0.35	1418	683	7751	0.85	0.33	11,014,267	4,465
	Google	0.36	4454	1914	7217	0.62	0.30	24,699,923	12,221
	Google_conf	0.40	4345	1217	6484	0.60	0.23	20,950,198	10,381
	OSM	**0.56**	3996	1099	5537	0.58	**0.22**	13,030,854	7,373
	OMF	0.50	4374	1540	5403	**0.55**	0.26	18,798,959	9,243
Ph	Bing	0.29	1921	1455	9718	0.83	0.43	17,421,764	8,227
	Google	0.29	4069	3632	9445	**0.70**	0.47	36,439,308	15,371
	Google_conf	**0.65**	3081	2089	8785	0.74	0.40	29,371,231	10,205
	OSM	0.44	1920	575	8998	0.82	**0.23**	10,984,408	4,885

in the Philippines, indicating a substantial proportion of predicted buildings were inaccurately identified. The FNR was also significantly high, ranging from 0.55 (OMF) to 0.85 (Bing) in Tanzania and from 0.70 (Google) to 0.83 (Bing). Notably, there was a clear difference in the FNR of OSM between Tanzania (0.58) and the Philippines (0.82), likely due to Tanzania's higher urban tile coverage in OSM compared to the more rural tiles in the Philippines, where OSM had no coverage (resulting in an FNR of 1 in those areas).

In both case studies, the Google dataset identified more buildings than the reference dataset. This discrepancy is partly attributable to built-up areas in larger settlements, where the reference dataset generalized multiple buildings as one, whereas the Google dataset detected individual structures. In addition, the Google dataset often included small buildings that were absent in the reference dataset and other datasets. The Google_conf dataset, with its confidence threshold, showed fewerFPs and FNs, thereby increasing accuracy. However, it is important to note that the confidence threshold also omitted some TPs, although this was less significant in Tanzania, where Google_conf exhibited a lower FDR and FNR compared to the regular Google dataset. The Bing dataset performed the worst, with the highest FNR and a relatively high FDR, showing difficulty in building detection and accurate geometric representation. OMF's performance was very similar to that of OSM in Tanzania, sometimes replicating the same buildings as Bing in areas without OSM coverage. In the Philippines, the two datasets are identical.

B. Equality of Opportunity and Predictive Parity

The analysis of equality of opportunity among sensitive attributes revealed variations across subgroups (see Table III). The OSM dataset in Tanzania displayed the lowest equality of opportunity, with significant variation within each variable, a trend not mirrored in the Philippines, likely due to OSM's limited coverage in many areas.

Of all the sensitive attributes, Rural_scale demonstrated the strongest impact on equality of opportunity. Bing had the highest overall equality of opportunity, but this was primarily due to consistently high FNR rates, underscoring the importance of not considering equality of opportunity in isolation, but rather in conjunction with actual FNR rates. Interestingly, the predictive parity rate was considerably lower than the equality of opportunity rate (see Table IV), suggesting greater variation between groups when focusing on FPs instead of FNs. OSM exhibited the lowest values for predictive parity in both Tanzania and the Philippines, with a negative correlation with RWI and a positive correlation with the Rural_scale variable being significant contributors to this trend. The results of the NBCs (see Fig. 1) displayed similar tendencies.

TABLE III
EQUALITY OF OPPORTUNITY RATES PER DATASET AND SENSITIVE VARIABLE (TZ = TANZANIA; PH = PHILIPPINES)

	Dataset	Pop_dens	RWI	Rural_scale	Bld_size
Tz	Bing	0.94	0.95	0.84	0.86
	Google	0.80	0.84	**0.74**	0.80
	Google_conf	0.80	0.83	**0.73**	**0.78**
	OSM	**0.46**	**0.36**	**0.43**	**0.75**
	OMF	**0.51**	**0.42**	**0.46**	**0.77**
Ph	Bing	0.91	0.93	**0.79**	0.92
	Google	0.93	0.86	**0.71**	0.83
	Google_conf	0.92	0.87	0.82	0.82
	OSM	0.80	0.89	**0.52**	0.83

TABLE IV
PREDICTIVE PARITY RATES PER DATASET AND SENSITIVE VARIABLE (TZ = TANZANIA; PH = PHILIPPINES)

	Dataset	Pop_dens	RWI	Rural_scale	Bld_size
Tz	Bing	**0.31**	**0.41**	**0.27**	**0.54**
	Google	**0.52**	**0.63**	**0.65**	**0.58**
	Google_conf	**0.45**	**0.58**	**0.30**	**0.50**
	OSM	**0.23**	**0.19**	**0.20**	**0.57**
	OMF	**0.27**	**0.27**	**0.22**	**0.57**
Ph	Bing	**0.67**	**0.72**	**0.40**	**0.86**
	Google	**0.74**	**0.71**	**0.69**	**0.65**
	Google_conf	**0.65**	**0.65**	**0.63**	**0.55**
	OSM	**0.45**	**0.19**	**0.10**	**0.40**

C. Visual Examples

Figs. 2 and 3 visually demonstrate examples of these patterns. Regarding Fig. 2, the OSM dataset lacked building data in this impoverished area, resulting in an FNR of 1. In contrast, Google's dataset identified nearly all buildings, a characteristic shared with Google_conf. However, Google_conf differs by excluding one incorrectly classified building present in Google's

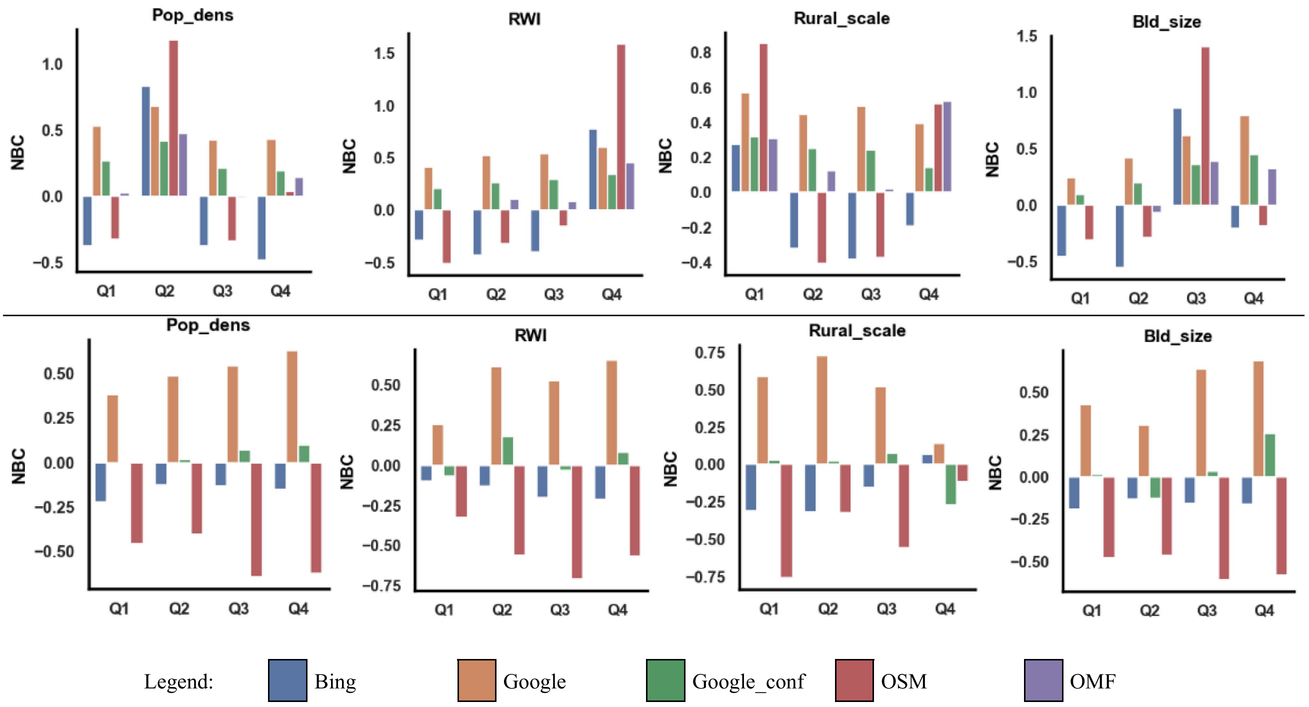

Fig. 1. NBC for Tanzania (row 1) and the Philippines (row 2), grouped by sensitive attributes.

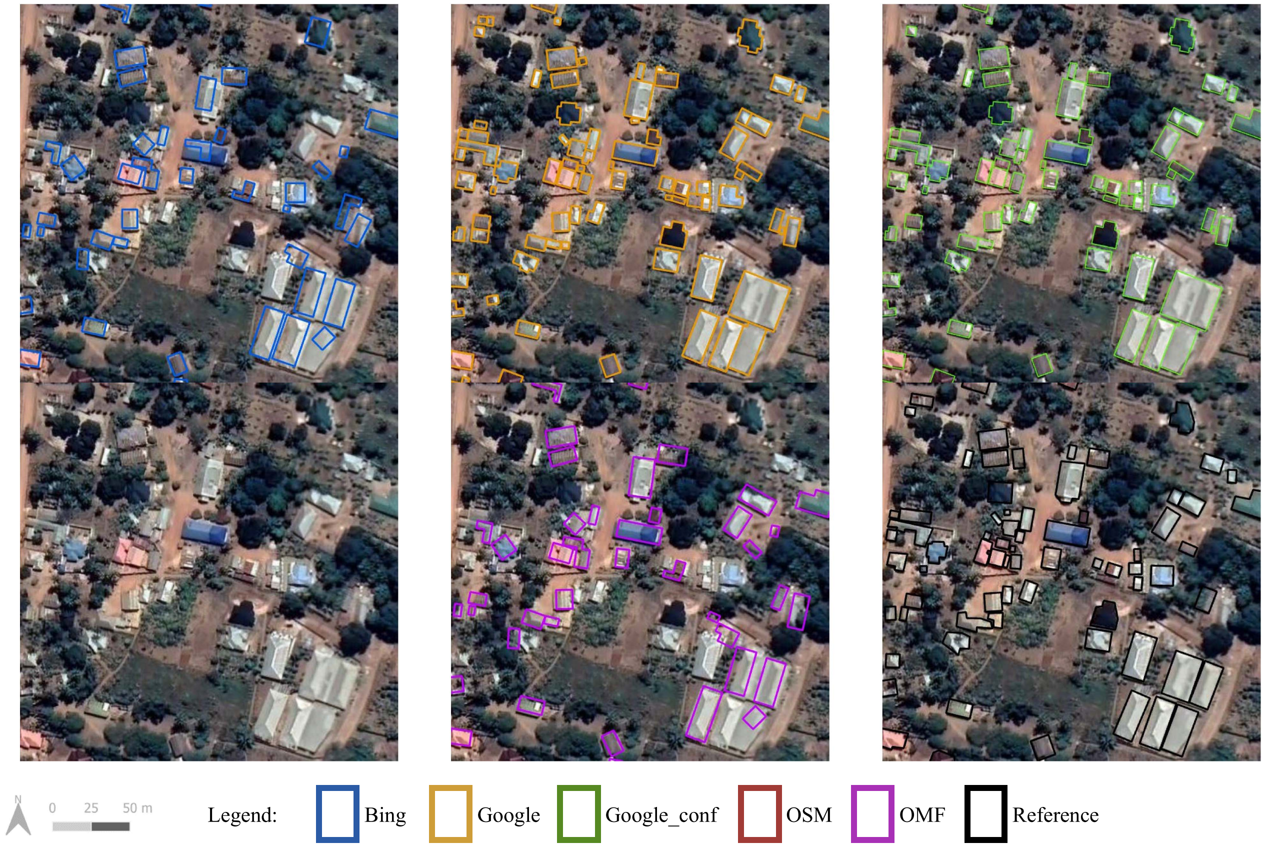

Fig. 2. Example of buildings delineated by the different datasets for a study area in Tanzania, which has a high FNR for OSM in low RWI areas compared to Google and Google_conf. Note that OSM (bottom right) does not map a single building. Tile statistics are as follows: RWI = 0.03 (Q1), Pop_dens (Q2) = 2.40, Rural = 6 (Q2), Bld_size = 141.8348 (Q3).

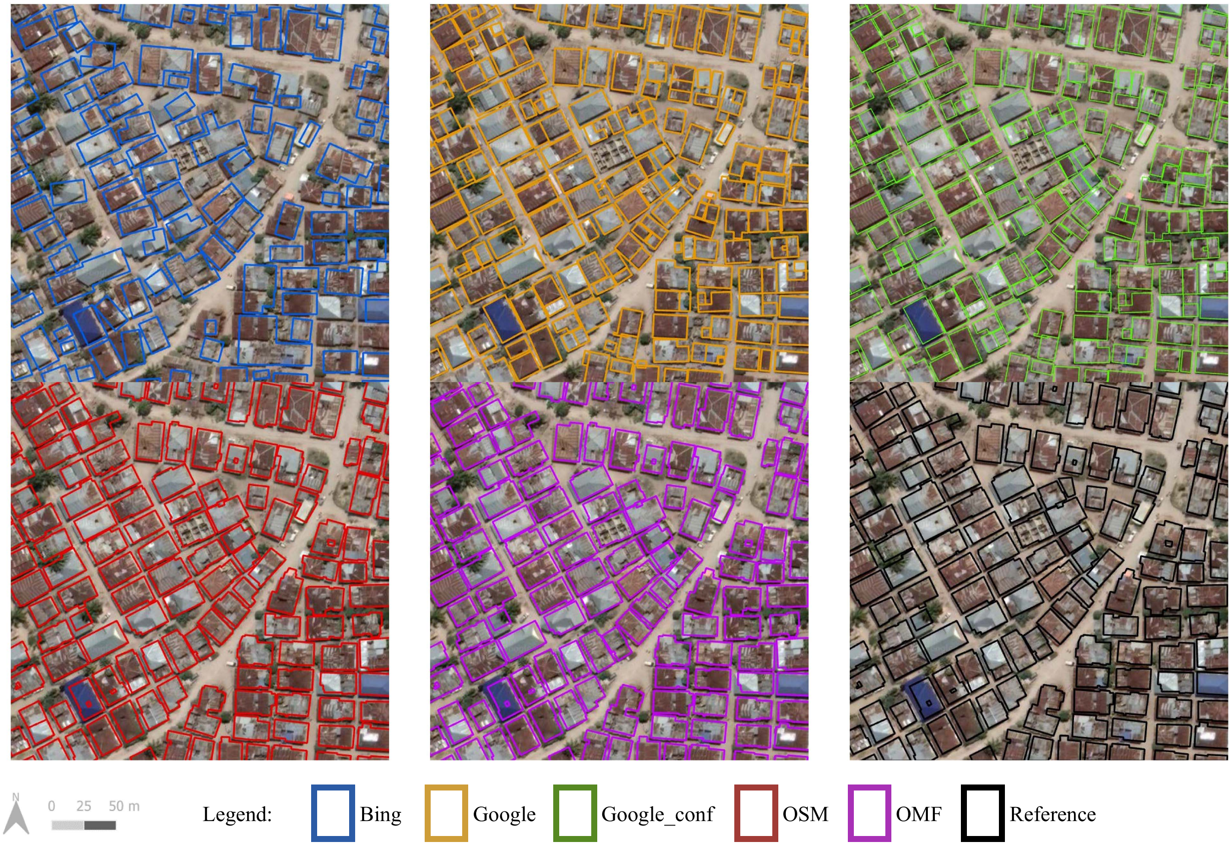

Fig. 3. Example of buildings delineated by the different datasets for a study area in Tanzania showing an opposite trend. This high RWI area has a lower FNR compared to Google and Google_conf. Tile statistics are as follows: RWI = 1.346 (Q4), Pop_dens = 23.995 (Q4), Rural_scale = 3 (Q1), Bld_size = 135.0579 (Q3).

data, while also missing another building. The Bing and OMF datasets exhibited significant similarities. Commonly, OMF mirrors OSM; however, in the absence of OSM data, it incorporates Bing's building data. Bing's dataset occasionally misses buildings and shows misplacement in some parts, whereas the OMF's dataset is slightly more accurate for building detection. The trends discussed are also evident in Fig. 4, specifically the RWI, Pop_dens, and Rural_scale variables in relation to FNR.

In the example of Fig. 3, which shows a region with high RWI, the OSM dataset excels by correctly identifying all buildings, resulting in an FNR of 0. Conversely, the Google dataset exhibits a substantially higher FNR. This increase is primarily due to the misclassification of several buildings, which are incorrectly identified as multiple structures rather than single units. Notably, the OMF dataset in this area is a perfect match with the OSM dataset. The Bing dataset, however, faces challenges, including misplaced buildings and errors in classifying multiple buildings as a single entity. These observations are supported by Fig. 4, which not only reflects trends in the RWI but also demonstrates similar patterns for the Pop_dens and Rural_scale variables.

V. Discussion

Our study was limited to assessing the algorithmic biases of building delineation datasets. However, the representation bias is an important factor. Another limitation is that we could not investigate the performance of building delineation datasets as a function of time. Often, a clear timestamp is not available, and usually, only a building delineation dataset based on the most recent imagery is available. For some applications, such as rapid urbanization, it would be best to have a building delineation dataset of the same moment in time as the disaster event for which DRM modeling is taking place.

This method of auditing for biases requires ground-truth data; in this case, true building outlines. Such a dataset of building outlines is not available (which is why it was necessary to consider using global building footprint datasets in the first place). Manual labeling was conducted over the selected areas for the purposes of this study. However, this labeling is time-consuming. Therefore, although investigating whether these trends in the accuracy of building footprint datasets across the globe would be very interesting from a practitioner's perspective, it would require extensive labeling efforts beyond the scope of this current study. However, this study provided a clear description of the methods and workflow that can be employed to conduct global validation.

Finally, it seems that one of the greatest challenges of auditing geospatial datasets for biases may be to obtain adequate data with sensitive attributes to perform audits. For example, RWI is computed at a global level and provided at a grid of 2.4 km. Therefore,

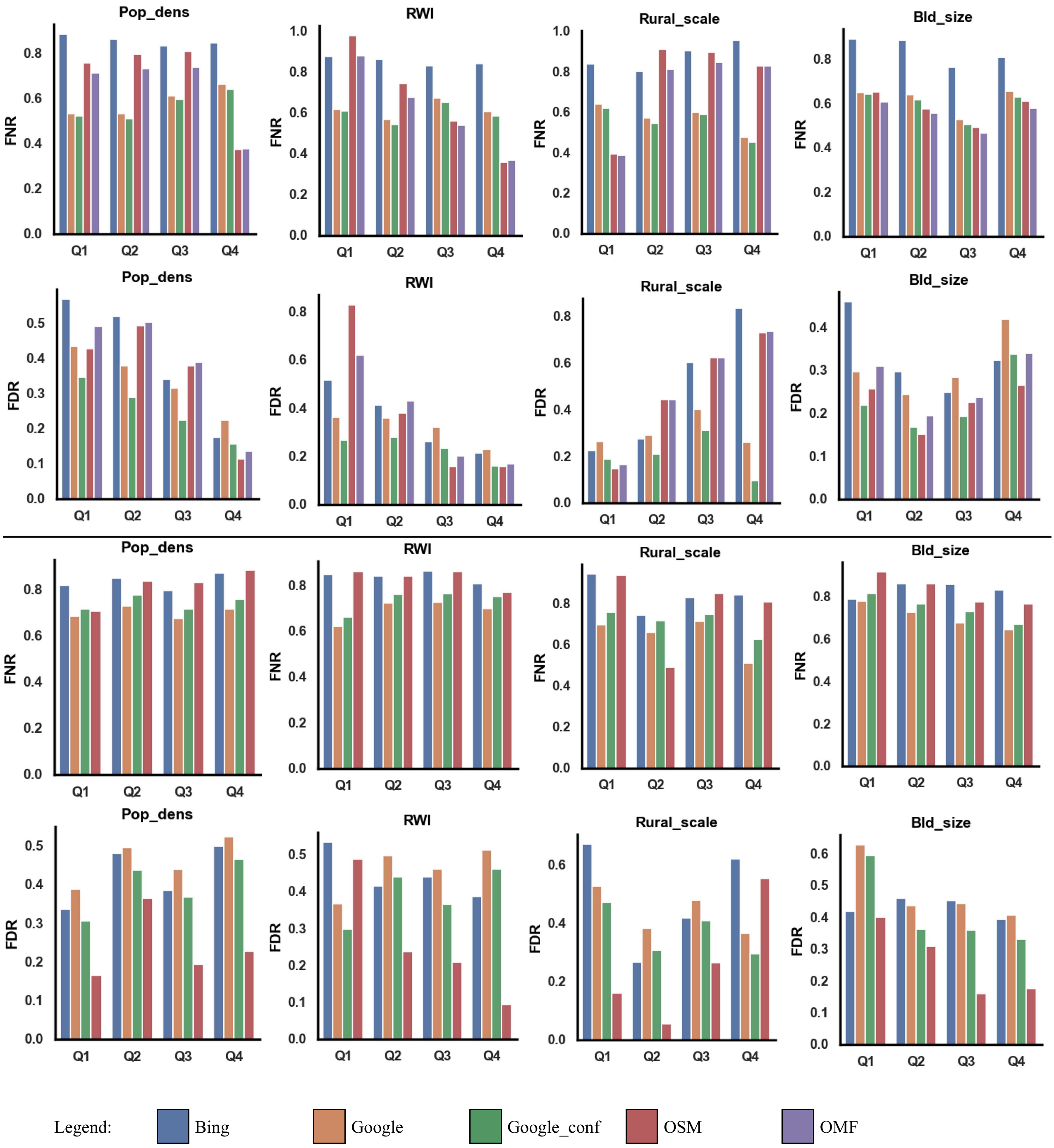

Fig. 4. Distribution plots for FNR and FDR for Tanzania (rows 1 and 2) and the Philippines (rows 3 and 4), grouped by sensitive attributes.

it may not be sufficiently nuanced to audit whether these algorithms perform equally well in deprived neighborhoods within larger cities. Correlations between sensitive attributes (e.g., RWI and Rural_scale) may further complicate the interpretation of the sensitivity analysis results. Indeed, some sensitive attributes (such as RWI) are determined using machine learning, and one must be careful that the building datasets to be audited are not utilized in the calculation of the sensitive attributes used for auditing. Ideally, sensitive attribute data can be obtained from independent sources such as household survey data. However, such data are often aggregated at the administrative unit level and are therefore not available at a sufficient spatial scale for the purposes of this study.

VI. Conclusion

This study analyzed the accuracy of four open-building delineation datasets and their biases across several sensitive attributes for two case studies. The Google and Microsoft building delineation datasets showed lower accuracy metrics for the two selected study areas (the Philippines and Tanzania) than the values officially reported at a global level. In addition, there were a substantial number of FNs across all datasets. Among the datasets tested, Google_conf and OSM had the best FNR and FDRs for Tanzania and the Philippines. An analysis of the distribution of FNR and FDR rates emphasizes that OSM has the greatest variation in the case study areas, with high accuracies in targeted areas but no buildings mapped in others. Regarding sensitive attributes, predictive parity rates were low across the board. Rural_scale was the most sensitive attribute with the lowest equality of opportunity for all datasets. OSM and OMF also displayed a low equality of opportunity values for the other three sensitive attributes (Pop_dens, RWI, and Bld_size). This study clearly indicates that there are biases in the accuracies of building datasets according to sensitive attributes, although the trends of these biases are not always consistent across the datasets and study areas.

The presence of these FNs means that, in DRM interventions, some buildings or households could potentially not receive support, whereas they would need it. Conversely, FPs could indicate that some will receive support when not actually needed. As distributing aid among vulnerable populations is considered a "no regrets" case in this context, the former is considered more worrying than the latter. Therefore, this study focuses on the equality of opportunity. When auditing other geospatial workflows for biases, it is important to consider the context when selecting which of the many fairness metrics to use.

Future research will investigate the sensitivity of specific AI models, such as convolutional neural networks for automated damage assessments [60] and XGBoost for predictive models [42] for the performance and biases of the input datasets. This also requires examining the aggregation bias, as some effects of the biases might average out if the models operate at a higher spatial resolution. For both automated damage assessments and predictive modeling, transferability is important, as new disaster events will usually not occur in the same feature space as the training and test data [61], [62]. Therefore, further research will consider more comprehensive audits by comparing more study areas to determine whether transferable and more generalized trends can be found. Additional research will also consider the effects of intersectional groups [63]. Some analyses were conducted in the current study, but it was concluded that the number of data points for the intersectional groups was too low to be conclusive.

Overall, our research is an important stepping stone towards auditing geospatial workflows for fairness and bias, especially in low- to middle-income countries where there is often a lack of local data. Quantifying the performance of the global dataset and its biases makes it possible to account for the equity of disaster risk management interventions. Biases in geospatial datasets emerge at various stages in automatic workflows, from the definition of the data model (i.e., what is a building and what is not), to the sampling of training data and model training and evaluation [9]. Bias mitigation measures can be taken at each stage in this loop. Still, methods to quantify datasets for biases, such as those presented here, can help practitioners understand the limitations in the data and adjust how the data is used in order to take possible inequalities into account when utilizing it for decision-making.

Acknowledgment

The authors would like to thank A. Teklesadik for the discussions and insights during the preparation of this manuscript. Map data copyrighted OSM contributors and available from https://www.openstreetmap.org.

References

[1] R. Soden, D. Lallemant, M. Kalirai, C. Liu, D. Wagenaar, and S. Jit, "The importance of accounting for equity in disaster risk models," *Commun. Earth Environ.*, vol. 4, no. 1, Oct. 2023, Art. no. 386, doi: 10.1038/s43247-023-01039-2.

[2] Red Cross, "Typhoon-impact-based-forecasting-model - github," 2021. Accessed: Jul. 15, 2023. [Online]. Available: https://github.com/rodekruis/Typhoon-Impact-based-forecasting-model

[3] M. Leidig and R. M. Teeuw, "Quantifying and mapping global data poverty," *PLoS One*, vol. 10, no. 11, Nov. 2015, Art. no. e0142076, doi: 10.1371/journal.pone.0142076.

[4] E. Haak, J. Ubacht, M. Van den Homberg, S. Cunningham, and B. Van den Walle, "A framework for strengthening data ecosystems to serve humanitarian purposes," in *Proc. 19th Annu. Int. Conf. Digit. Government Res., Governance Data Age*, May 2018, pp. 1–9. doi: 10.1145/3209281.3209326.

[5] Google, "Open buildings," 2021. Accessed: Sep. 01, 2022. [Online]. Available: https://sites.research.google/open-buildings/

[6] Microsoft, "Uganda-Tanzania-building-footprints," 2020. Accessed: Sep. 01, 2022. [Online]. Available: https://github.com/microsoft/Uganda-Tanzania-Building-Footprints

[7] Microsoft, "IdMyPhBuildingFootprints GitHub repo," 2022. Accessed: Dec. 20, 2023. [Online]. Available: https://github.com/microsoft/IdMyPhBuildingFootprints

[8] Overture Maps Foundation, "Overture maps foundation," 2023. Accessed: Dec. 01, 2023. [Online]. Available: https://overturemaps.org/

[9] H. Suresh and J. Guttag, "A framework for understanding sources of harm throughout the machine learning life cycle," in *Equity and Access in Algorithms, Mechanisms, and Optimization*, New York, NY, USA: ACM, Oct. 2021, pp. 1–9, doi: 10.1145/3465416.3483305.

[10] M. Favaretto, E. De Clercq, and B. S. Elger, "Big Data and discrimination: Perils, promises and solutions. A systematic review," *J. Big Data*, vol. 6, no. 1, Dec. 2019, Art. no. 12, doi: 10.1186/s40537-019-0177-4.

[11] J. G. Cavazos, P. J. Phillips, C. D. Castillo, and A. J. O'Toole, "Accuracy comparison across face recognition algorithms: Where are we on measuring race bias?," *IEEE Trans. Biom. Behav. Identity Sci.*, vol. 3, no. 1, pp. 101–111, Jan. 2021, doi: 10.1109/TBIOM.2020.3027269.

[12] Z. Wang et al., "Towards fairness in visual recognition: Effective strategies for bias mitigation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 8916–8925, doi: 10.1109/CVPR42600.2020.00894.

[13] J. Banasik, J. Crook, and L. Thomas, "Sample selection bias in credit scoring models," *J. Oper. Res. Soc.*, vol. 54, no. 8, pp. 822–832, Aug. 2003, doi: 10.1057/palgrave.jors.2601578.

[14] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-aware classifier with prejudice remover regularizer," in *Proc. Mach. Learn. Knowl. Discov. Databases: Eur. Conf.*, 2012, pp. 35–50, doi: 10.1007/978-3-642-33486-3_3.

[15] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big Data*, vol. 5, no. 2, pp. 153–163, Jun. 2017, doi: 10.1089/big.2016.0047.

[16] G. Boo et al., "High-resolution population estimation using household survey data and building footprints," *Nat. Commun.*, vol. 13, no. 1, Mar. 2022, Art. no. 1330, doi: 10.1038/s41467-022-29094-x.

[17] C. Robinson et al., “Rapid building damage assessment workflow: An implementation for the 2023 Rolling Fork, Mississippi tornado event,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 3760–3764. Accessed: Apr. 25, 2024. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2023W/AIHADR/html/Robinson_Rapid_Building_Damage_Assessment_Workflow_An_Implementation_for_the_2023_ICCVW_2023_paper.html

[18] K. Tzavella, A. Skopeliti, and A. Fekete, “Volunteered geographic information use in crisis, emergency and disaster management: A scoping review and a web atlas,” *Geo-spatial Inf. Sci.*, vol. 27, no. 2, pp. 423–454, Mar. 2024, doi: 10.1080/10095020.2022.2139642.

[19] M. Kooshki Forooshani et al., “Towards a global impact-based forecasting model for tropical cyclones,” *Natural Hazards Earth System Sci.*, vol. 24, no. 1, pp. 309–329, Feb. 2024, doi: 10.5194/nhess-24-309-2024.

[20] R. Marty and A. Duhaut, “Global poverty estimation using private and public sector big data sources,” *Sci. Rep.*, vol. 14, no. 1, Feb. 2024, Art. no. 3160, doi: 10.1038/s41598-023-49564-6.

[21] A. Abascal et al., “AI perceives like a local: Predicting citizen deprivation perception using satellite imagery,” *npj Urban Sustainability*, vol. 4, no. 1, Mar. 2024, Art. no. 20, doi: 10.1038/s42949-024-00156-x.

[22] B. Herfort, S. Lautenbach, J. Porto de Albuquerque, J. Anderson, and A. Zipf, “A spatio-temporal analysis investigating completeness and inequalities of global urban building data in OpenStreetMap,” *Nat. Commun.*, vol. 14, no. 1, Jul. 2023, Art. no. 3985, doi: 10.1038/s41467-023-39698-6.

[23] H. R. Chamberlain, E. Darin, W. A. Adewole, W. C. Jochem, A. N. Lazar, and A. J. Tatem, “Building footprint data for countries in Africa: To what extent are existing data products comparable?,” *Comput. Environ. Urban Syst.*, vol. 110, Jun. 2024, Art. no. 102104, doi: 10.1016/j.compenvurbsys.2024.102104.

[24] M. Kuffer, M. Owusu, L. Oliveira, R. Sliuzas, and F. van Rijn, “The missing millions in maps: Exploring causes of uncertainties in global gridded population datasets,” *ISPRS Int. J. Geoinf.*, vol. 11, no. 7, Jul. 2022, Art. no. 403, doi: 10.3390/ijgi11070403.

[25] UNESCO, “Recommendation on the ethics of artificial intelligence,” France, 2021. Accessed: Nov. 03, 2022. [Online]. Available: https://unesdoc.unesco.org/ark:/48223/pf0000380455

[26] European Commission, “Proposal for a regulation of the European Parliament and of the council laying down harmonised rules on ARtificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts,” Apr. 2021. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

[27] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning*, 2019. Accessed: Aug. 07, 2023. [Online]. Available: http://www.fairmlbook.org

[28] A. Castelnovo, R. Crupi, G. Greco, D. Regoli, I. G. Penco, and A. C. Cosentini, “A clarification of the nuances in the fairness metrics landscape,” *Sci. Rep.*, vol. 12, no. 1, Mar. 2022, Art. no. 4209, doi: 10.1038/s41598-022-07939-1.

[29] A. N. Carey and X. Wu, “The statistical fairness field guide: Perspectives from social and formal sciences,” *AI Ethics*, vol. 3, no. 1, pp. 1–23, Feb. 2023, doi: 10.1007/s43681-022-00183-3.

[30] A. Chouldechova and A. Roth, “A snapshot of the frontiers of fairness in machine learning,” *Commun. ACM*, vol. 63, no. 5, pp. 82–89, 2020, doi: 10.1145/3376898.

[31] M. Hardt, E. Price, and N. Srebro, “Equality of opportunity in supervised learning,” in *Proc. Adv. Neural Inf. Process. Syst.*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., Curran Associates, 2016, pp. 3315–3323. Accessed: Jul. 08, 2023. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf

[32] A. Chouldechova, “Fair prediction with disparate impact: A study of bias in recidivism prediction instruments,” *Big Data*, vol. 5, no. 2, pp. 153–163, Jun. 2017, doi: 10.1089/big.2016.0047.

[33] J. Kleinberg, S. Mullainathan, and M. Raghavan, “Inherent trade-offs in the fair determination of risk scores,” Sep. 2016.

[34] Center for Data Science and Public Policy, “Aequitas – Bias and fairness audit – Fairness Tree,” 2018. Accessed: Aug. 07, 2023. [Online]. Available: http://www.datasciencepublicpolicy.org/our-work/tools-guides/aequitas/

[35] J. Rawls, *Justice as Fairness: A Restatement*. Cambridge, MA, USA: Harvard Univ. Press, 2001.

[36] A. Ghosh, L. Genuit, and M. Reagan, “Characterizing intersectional group fairness with worst-case comparisons,” in *Proc. Mach. Learn. Res.*, 2021, pp. 22–34.

[37] H. Weerts, M. Dudík, R. Edgar, A. Jalali, R. Lutz, and M. Madaio, “Fairlearn: Assessing and improving fairness of AI systems,” *J. Mach. Learn. Res.*, vol. 24, no. 257, pp. 1–8, 2023. [Online]. Available: http://jmlr.org/papers/v24/23-0389.html

[38] World Bank, “DataBank World Development indicators,” 2023. Accessed: Apr. 25, 2024. [Online]. Available: https://data.worldbank.org/indicator/SP.URB.TOTL.IN.ZS?locations=TZ-PH

[39] M. Mukim and A. Chunet, “Mapping urban poverty from space,” World Bank Blogs, 2022. Accessed: Dec. 02, 2023. [Online]. Available: https://blogs.worldbank.org/sustainablecities/mapping-urban-poverty-space

[40] S. Ferguson, M. Van Ledden, S. L. Rubinyi, A. Campos Garcia, and T. M. Doeffinger, “Urban Flood Risk Handbook: Assessing risk and identifying interventions,” Washington D.C., 2023. Accessed: Dec. 02, 2023. [Online]. Available: http://documents.worldbank.org/curated/en/099080123151036528/P1744620efe1180a20bc1b0ce287e74ff91

[41] M. J. C. Van den Homberg, C. M. Gevaert, and Y. Georgiadou, “The changing face of accountability in humanitarianism: Using artificial intelligence for anticipatory action,” *Politics Governance*, vol. 8, no. 4, pp. 456–467, Dec. 2020, doi: 10.17645/pag.v8i4.3158.

[42] M. Kooshki et al., “Towards a global impact-based forecasting model for tropical cyclones,” EGUsphere, 2023, doi: 10.5194/egusphere-2023-2205.

[43] OpenStreetMap, “OpenStreetMap,” 2023. Accessed: Aug. 01, 2023. [Online]. Available: https://www.openstreetmap.org/

[44] HOTOSM, “Tanzania,” HOTOSM - Where we work, 2023. Accessed: Dec. 01, 2023. [Online]. Available: https://www.hotosm.org/where-we-work/tanzania/

[45] Ramani Huria : The atlas of flood resilience in Dar es Salaam (English), Washington, D.C. : World Bank Group, 2018. [Online]. Available: http://documents.worldbank.org/curated/en/200421524092301920/Ramani-Huria-the-atlas-of-flood-resilience-in-Dar-es-Salaam

[46] HOTOSM, “HOTOSM Tanzania buildings (OpenStreetMap Export).” Accessed: Dec. 18, 2023. [Online]. Available: https://data.humdata.org/dataset/hotosm_tza_buildings

[47] HOTOSM, “HOT tasking Manager,” 2023. Accessed: Dec. 20, 2023. [Online]. Available: https://tasks.hotosm.org/explore?text=Philippines

[48] W. Sirko, “Continental-scale building detection from high resolution satellite imagery,” Jun. 2021.

[49] Overture Maps Foundation, “OvertureMaps Github repo.” Accessed: Oct. 01, 2023. [Online]. Available: https://github.com/OvertureMaps/data#data-release-feedback

[50] Facebook, “Relative Wealth Index,” 2022. Accessed: Nov. 02, 2023. [Online]. Available: https://dataforgood.facebook.com/dfg/tools/relative-wealth-index#methodology

[51] HDX, “Relative Wealth Index,” 2021. Accessed: Nov. 02, 2023. [Online]. Available: https://data.humdata.org/dataset/relative-wealth-index?

[52] Facebook, “High resolution population density maps,” 2023. Accessed: Nov. 10, 2023. [Online]. Available: https://dataforgood.facebook.com/dfg/tools/high-resolution-population-density-maps

[53] HDX, “High resolution density maps & demographic estimates,” 2019. Accessed: Nov. 10, 2023. [Online]. Available: https://data.humdata.org/dataset/Philippines-high-resolution-population-densitymaps-demographic-estimates?

[54] FAO, “Global urban rural catchment areas (URCA) grid - 2021,” 2021. Accessed: Dec. 02, 2023. [Online]. Available: https://data.apps.fao.org/map/catalog/srv/api/records/9dc31512-a438-4b59-acfd-72830fbd6943/formatters/xml?approved=true

[55] DLR, “World settlement footprint (WSF) 2019 - Sentinel-1/2 - global,” 2019. Accessed: Dec. 02, 2023. [Online]. Available: https://download.geoservice.dlr.de/WSF2019/

[56] M. Kuffer, K. Pfeffer, and R. Sliuzas, “Slums from space—15 years of slum mapping using remote sensing,” *Remote Sens (Basel)*, vol. 8, no. 6, May 2016, Art. no. 455, doi: 10.3390/rs8060455.

[57] W. Zhao, C. Persello, and A. Stein, “Building outline delineation: From aerial images to polygons with an improved end-to-end learning framework,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 175, pp. 119–131, May 2021, doi: 10.1016/j.isprsjprs.2021.02.014.

[58] S. P. Mohanty et al., “Deep Learning for understanding satellite imagery: An experimental survey,” *Front. Artif. Intell.*, vol. 3, 2020, Art. no. 534696, doi: 10.3389/frai.2020.534696.

[59] P. Saleiro et al., “Aequitas: A bias and fairness audit toolkit,” Nov. 2018. [Online]. Available: http://arxiv.org/abs/1811.05577

[60] B. Herfort, S. Lautenbach, J. Porto de Albuquerque, J. Anderson, and A. Zipf, “A spatio-temporal analysis investigating completeness and inequalities of global urban building data in OpenStreetMap,” *Nat. Commun.*, vol. 14, no. 1, Jul. 2023, Art. no. 3985, doi: 10.1038/s41467-023-39698-6.

[61] T. Valentijn, J. Margutti, M. van den Homberg, and J. Laaksonen, “Multi-hazard and spatial transferability of a CNN for automated building damage assessment,” *Remote Sens (Basel)*, vol. 12, no. 17, Sep. 2020, Art. no. 2839, doi: 10.3390/rs12172839.

[62] D. Wagenaar et al., “Improved transferability of data-driven damage models through sample selection bias correction,” *Risk Anal.*, vol. 41, no. 1, pp. 37–55, Jan. 2021, doi: 10.1111/risa.13575.

[63] S. J. Pan and Q. Yang, “A survey on transfer learning,” *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010, doi: 10.1109/TKDE.2009.191.

[64] F. Yang, M. Cisse, and S. Koyejo, “Fairness with overlapping groups,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 4067–4078.

Caroline M. Gevaert received the B.Sc. degree in international land and water management from Wageningen University, Wageningen, The Netherlands, in 2011, the M.Sc. degree in remote sensing from the University of Valencia, Valencia, Spain, in 2013, and the M.Sc. degree in geographical information science from Lund University, Lund, Sweden, in 2014.

She is an Associate Professor with the Faculty Geo-information Science and Earth Observation (ITC), University of Twente, Enschede, The Netherlands, and an International Consultant for the World Bank. Her research investigated the use of satellite and drone imagery and machine learning to support development projects in lower- to middle-income countries.

Dr. Gevaert was the recipient of the Cum Laude distinction for the Ph.D. dissertation on mapping informal settlements with drones and the NCG J.M. Tienstra research prizes. She was a member of the Dutch Young Academy of Scientists. More recently, she has received a prestigious personal research grant from the Dutch Research Council to investigate how to make AI in geospatial workflows fair and explainable.

Thomas Buunk received the bachelor’s degree in beta-gamma, which focuses on advanced geospatial technologies and applications, from the University of Amsterdam, Amsterdam, The Netherlands, in 2021, and the Master’s degree in geographical information management and applications, a program collaboratively offered by Delft University of Technology, Delft, The Netherlands, University of Twente, Enschede, Netherlands, Utrecht University, Utrecht, The Netherlands, and Wageningen University & Research, Wageningen, The Netherlands, in 2023.

During the bachelor’s degree, he engaged in an interdisciplinary curriculum combining natural sciences, economics, and social sciences, with a major in human geography. In the last year of his Master’s thesis, he conducted a thesis project on crop stress analysis within precision viticulture. The research aligned with his fascination for sustainable environmental management and culminated in scientific publication. For the final stage of his Master’s degree, he completed an internship at 510, an initiative of The Netherlands Red Cross. He gained experience in applying geospatial technologies in a humanitarian context by contributing to this project. He is committed to contributing to the field’s growth and the development of innovative, data-driven solutions for societal benefit.

Marc J.C. van den Homberg received the Ph.D. degree in physics from Delft University of Technology, Delft, The Netherlands, in 1998, the MBA degree from Rotterdam School of Management, Rotterdam, The Netherlands, in 2005, and a disaster management certificate from IFRC/Tata Institute of Social Sciences.

He is the Scientific Lead of 510, an initiative of The Netherlands Red Cross. He is a Professor on the Princess Margriet Chair “Data for disaster resilience” at ITC/Faculty of Geo-Information Science and Earth Observation. His research interests include improving preparedness and response to both natural hazards and complex emergencies through data-driven risk assessments, impact-based forecasting, information management, and disaster risk governance. In particular, he focused on enhancing the quality of risk and impact data, triggering models for anticipatory action (preparedness), and selecting, monitoring, and building evidence for nature-based solutions (prevention).

He is a member of the IFRCs Early Action Protocol Validation Committee Advisory and the Advisory Board of WMO’s High-Impact Weather Project (HIWeather). He has helped shape the growth of 510 from its inception in 2016 by developing and implementing an applied research agenda tailored to the needs of the Red Cross Red Crescent Movement and establishing close collaborations with universities and knowledge institutes. This has led to evidence and peer review of products and services–introduced by The Netherlands Red Cross–aiming to improve the quality, speed, and cost-effectiveness of humanitarian action. Before joining 510, he led and cofounded the ICT for Development (ICT4D) team within TNO, The Netherlands Research and Technology Organization with which he implemented pro-poor ICT innovations.