







Digital Injustice: A Case Study of Land Use Classification Using Multisource Data in Nairobi, Kenya

Wenlan Zhang   

Centre for Advanced Spatial Analysis, University College London, UK

Chen Zhong¹   

Centre for Advanced Spatial Analysis, University College London, UK

Faith Taylor   

Department of Geography, King's College London, UK

Centre for Advanced Spatial Analysis, University College London, UK

Abstract

The utilisation of big data has emerged as a critical instrument for land use classification and decision-making processes due to its high spatiotemporal accuracy and ability to diminish manual data collection. However, the reliability and feasibility of big data are still controversial, the most important of which is whether it can represent the whole population with justice. The present study incorporates multiple data sources to facilitate land use classification while proving the existence of data bias caused digital injustice. Using Nairobi, Kenya, as a case study and employing a random forest classifier as a benchmark, this research combines satellite imagery, night-time light images, building footprint, Twitter posts, and street view images. The findings of the land use classification also disclose the presence of data bias resulting from the inadequate coverage of social media and street view data, potentially contributing to injustice in big data-informed decision-making. Strategies to mitigate such digital injustice situations are briefly discussed here, and more in-depth exploration remains for future work.

2012 ACM Subject Classification Applied computing → Environmental sciences

Keywords and phrases Data bias, Digital injustice, Multi-source sensor data, Land use classification, Random forest classifier

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.94

Category Short Paper

Funding *Chen Zhong*: The research has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 949670).

1 Introduction

Land use classification is an essential part of resource distribution such as conducting infrastructure upgrading projects and services provision activities. It is widely accepted to classify land use types using remote sensing data with census, survey, or interview [3]. Despite providing high-accuracy information, the traditional classification methods have common disadvantages of being labour-intensive, time-consuming, low spatial resolution and requiring substantial financial resources, which create barriers for the Global South countries to apply [5]. It is crucial to offer cost-effective and easily accessible methods for land use classification to decision-makers in the Global South. This would enable underprivileged countries to receive timely and precise information required for emergency assistance provision.

¹ Corresponding author



© Wenlan Zhang, Chen Zhong, and Faith Taylor;
licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 94; pp. 94:1–94:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Unlike traditional survey-based data, big data - referring to sensor-collected automatic data - has gradually become a low-cost, timely, cost-efficient supplement to the traditional data sources in the Global South countries [6]. As a by-product of advancing technology and digitalisation, new data sources (e.g., social media, street view image) are generally collected by Internet of Things sensors and smart devices in the form of social media data, street view data, and remote sensing data [9]. The datasets can contain various information such as geo-referenced text, images, and GPS signals. This information can be used to analyse people's social activity patterns, and even hence infer the land use types.

However, it is estimated that 37% of the global population remains to have restricted or no access to the internet, and the disconnected proportion is unsurprisingly high in Global South countries. Those with no access to smart devices or the internet are called 'digitally invisible' since they have less opportunity to generate data that could influence policy or benefit from data-informed analysis [2]. This data-caused discrimination, together with visibility and engagement with technology, was concluded as a data justice challenge by Prof. Linnet Taylor [8]. Data bias and the impact of digital injustice have created an obstacle to the application of big data. However, limited research has been conducted to verify digital injustice and to propose effective strategies for its mitigation. Therefore, this research aims to identify instances of digital injustice by performing a land use classification using multi-source publicly available data, with a case study of Nairobi, Kenya. The question of who constitutes the digitally invisible groups and where they reside remains an unresolved issue for future work.

2 Study Materials and Methodology

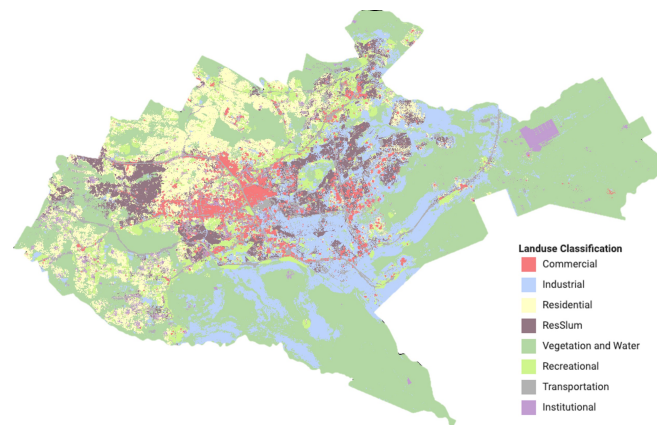
The case study city of Nairobi is the capital city of Kenya, which has been the economic centre of East Africa [4], experiencing overwhelming population growth and informal settlements expansion. These informal settlement areas accommodate more than 60% of the total population while occupying less than 5% of the city's residential land area [1]. Due to the rapid pace of development, the land use and extent of informal settlements can change significantly within a short period of time. Therefore, frequently updated land use data would be beneficial for local decision-makers.

The 2010 land use map shapefile of Nairobi, Kenya, created by Columbia University's Center for Sustainable Urban Development and obtained from the World Bank Data Catalog, served as the training dataset. However, due to the prolonged interval since its release and the swift pace of urban development in Kenya, various modifications were implemented based on field investigations and comparisons using Google Maps. The initial dataset encompassed 13 categories, which were subsequently condensed to 8 categories in accordance with the Nairobi land use policy, namely, commercial, industrial, residential, informal settlements, vegetation, water, recreational, transportation, and institutional.

Multiple sources of open sensor data were employed to conduct the research, and the relevant information is summarised in the table 1.

■ **Table 1** Data source and feature.

| Data (abbr.) | Raster Data | | Vector Data | | |
|------------------------|--------------------------------------|-------------------------|-------------------------|-------------------------|--------------------------------|
| | Satellite images (R) | Night-time light (N) | Building footprints (O) | Social media posts (T) | Street view images (S) |
| Source | Sentinel-2 MSI | VIIRS-DNB | Google Open Building | Twitter posts | Mapillary |
| Information | Spatial resolution 10m | Spatial resolution 760m | Polygon | Text point | Image point |
| Feature selection | Bands, NDVI, NDWI, NDBI ² | Night Band | Building density | Tweet language and time | Object detected |
| Feature interpretation | Land physical char | Urban extent | Building char | Social activity | Sectional physical environment |



■ **Figure 1** Land use map with 8 categories.

The raster data was resampled to a unified spatial resolution of 30 meters to provide detailed information suitable for community and city-level analysis. However, this resolution was chosen primarily for illustration purposes, and the accuracy trend is expected to perform similarly across different spatial units. Twitter posts were categorized into three categories: working/school, leisure time, and home time, based on whether the post was made on a weekday or weekend and the time of the post. The content of the posts was analysed using language detection techniques. A panoramic segmentation of the street view images was conducted using Detectron2, a pre-trained object detection algorithm developed by Meta. The processed vector dataset was then rasterized to a 30-meter resolution to align with the remote sensing data.

In this study, the random forest was employed as the benchmark classifier for land use classification for illustration purposes, since it has been widely applied and considered to be the most effective method for land use and land cover classifier [7]. Random forest is a supervised learning technique, whereby the classification categories can be allocated from the training dataset. A sample of 1000 pixels was randomly selected from each category and split into training and test sets. The forest number was set to 200. It is worth noting, however, that the selection of the classifier does not constitute the primary objective of this research, and other classifiers could be utilised in lieu of random forest. Although the overall accuracy of different data combinations may differ, significant modifications to the ranking of overall accuracy are not anticipated.

3 Result and Discussion

3.1 Land use map

The predicted Nairobi land use map with a 30m spatial resolution is presented in Figure 1. Figure 2(a) illustrates the change in OA with different data combinations. The combination of all datasets achieved the highest overall accuracy of 71.57%. As hypothesised, the aggregation of multiple data sources significantly enhanced the effectiveness of the OA of land use classification. This trend is consistent across different spatial resolutions, as shown in Figure 2(b).

² NDVI: Normalised Difference Vegetation Index, NDWI: Normalised Difference Water Index, NDBI: Normalised Difference Build up Index

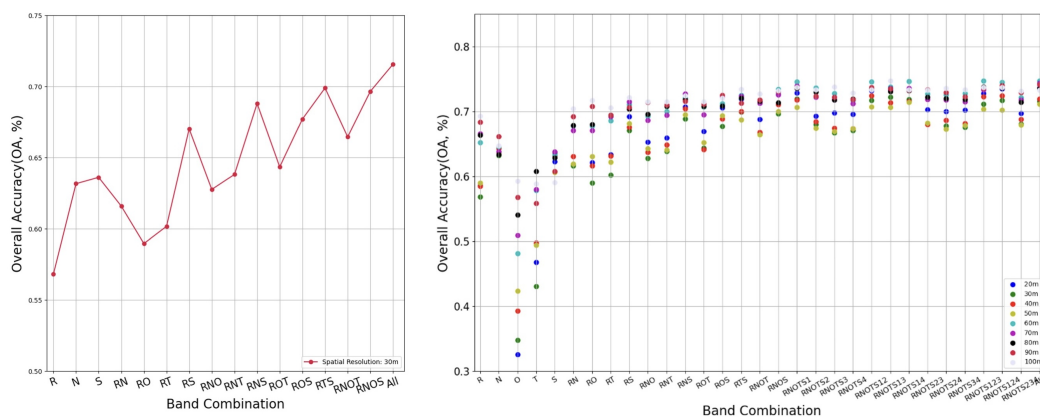


Figure 2 Land use classification (a) OA of 30m spatial resolution; (b) OA across spatial resolution.

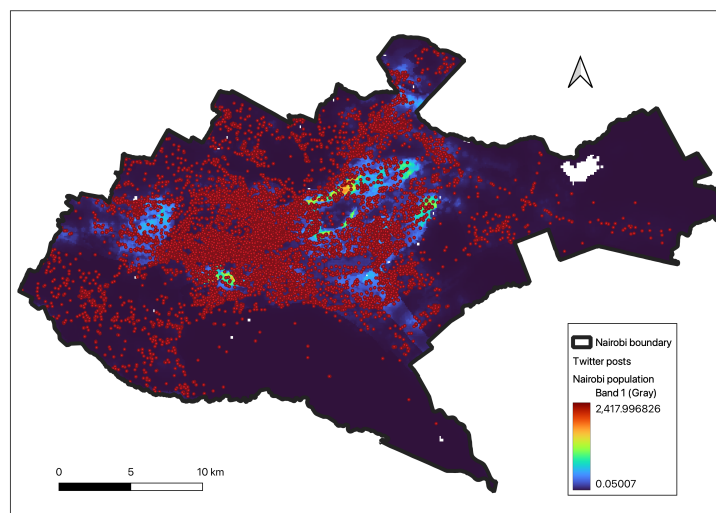
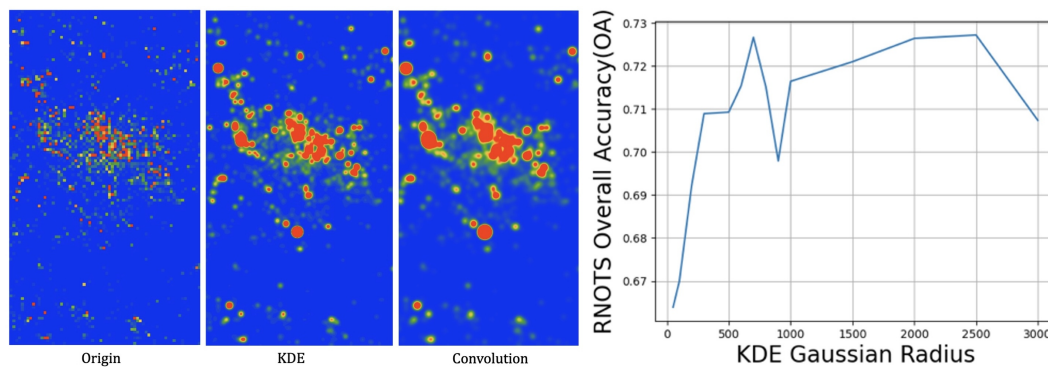


Figure 3 Twitter data biased spatial distribution.

This performance could be attributed to the fact that data aggregation allowed for a full range of information to be revealed. The satellite images, night-time light images, and building density reflect the physical features of the land. In addition, social and economic features, such as different languages (English, Swahili, or others) found in Twitter posts, can provide insight into the people’s education levels and social connectivity, as they may use English exclusively for professional and outreach activities in commercial areas. Industrial areas are among the most commonly used places for Swahili, which could contribute to increased accuracy. Moreover, the presence of umbrellas in street view images could be used to directly infer the presence of a commercial area, as it is a unique indication of a local roadside market. These findings provide further evidence of the importance of using mobility data to identify social and economic features.

3.2 Mitigating data injustice

The results also highlighted the existence of low-data areas in Nairobi, as can be seen in Figure 3. Some highly populated area (colour in red and yellow in the background), especially the informal settlements of Kibera and Mathare, was not covered by Twitter post. After



■ **Figure 4** KDE and convolution (a) Regional performance (b) OA performance.

dividing the city into a grid with a spatial resolution of 30 meters, there were a total of 792,534 grid cells. However, only 307,632 cells had valid Twitter posts with identified language, which accounted for 38.82% of the total area. People who live in places where no data is collected are digitally invisible groups. The existence of digitally invisible groups would reduce classification accuracy, and potentially lead to biased decision-making.

The possible reasons for this uneven data distribution were: (1) genuinely less populated areas: the urban outskirts contain underdeveloped bare land and agricultural land. (2) low internet or smart device penetration: as mentioned before the rural area would have lower smartphone access. (3) a preference for other social media platforms: according to research done by Kepios (Kemp, 2022), the social media preference ranking: Facebook (42.6%), LinkedIn > (12.4%) > Instagram (10.7%) > Snapchat (7.5%) > Twitter (5.8%).

Tobler's First Law of Geography suggests that neighbouring areas are more similar than distant ones. Based on which, we assume that increasing the impact of a single data point could potentially cover nearby no-data areas and amplify the voices of digitally invisible groups. This could be implemented by performing a kernel density estimation (KDE), followed by a Gaussian convolution, as shown in Figure 4(a). The land use classification accuracy with all bands (at a spatial resolution of 30 m) increased from 57.68% to 70.24%. This result proved our assumption that nearby land use can be inferred using single data points.

Determining the impact range of a single data point remains a critical question. To understand the effect of distance on land use classification accuracy, an optimisation of the parameter has been plotted as shown in Figure 4(b). The optimal performance distance for Twitter posts in Nairobi was approximately 700 meters, resulting in an OA of 72.72%. This finding suggests that land use types tend to remain consistent within a 700-meter radius in Nairobi. However, it should be noted that this approach only addresses digital injustices within a specific range. As the distance increases, land use categories may differ significantly, and thus, data gaps for large data-missing areas cannot be inferred. Therefore, designing surveys and interviews as supplementary data collection to visualise the digitally invisible groups for large data-missing areas would be beneficial.

4 Limitation and Future Work

This project is subject to certain limitations that need to be acknowledged. Firstly, the findings highlight the presence of data bias and digital injustice, along with a brief analysis of their spatial extent. However, the quantitative spatial coverage and representativeness of

the sensor data were not fully explored, which leaves open questions about the demographic, spatial, and temporal distribution of the digitally invisible population. Consequently, only limited mitigation approaches were provided, and no information was provided about who should be the target group for the small data collection. The unresolved inquiries also include whether big and small datasets can represent different social groups and whether performing data fusion can be implemented to mitigate digital injustice. These questions will be further explored in the next phase of our research.

The predicted land use map may not fully capture areas with multiple functions due to the relatively coarse 30m spatial resolution. This limitation is caused by the limited computing capacity of GEE. However, for city-level decision-making, a 30m resolution is generally sufficient. To overcome this limitation for more granular analyses, one can zoom in to a smaller area or switch to another server.

References

- 1 Stefanos Georganos, Angela Abascal, Monika Kuffer, Jiong Wang, Maxwell Owusu, Eléonore Wolff, and Sabine Vanhuysse. Is it all the same? mapping and characterizing deprived urban areas using worldview-3 superspectral imagery. a case study in nairobi, kenya. *Remote Sensing*, 13, December 2021. doi:10.3390/rs13244986.
- 2 Justin Longo, Evan Kuras, Holly Smith, David M. Hondula, and Erik Johnston. Technology use, exposure to natural hazards, and being digitally invisible: Implications for policy analytics. *Policy and Internet*, 9:76–108, March 2017. doi:10.1002/POI3.144.
- 3 Darius Phiri, Matamayo Simwanda, Serajis Salekin, Vincent R. Ryirenda, Yuji Murayama, Manjula Ranagalage, Nadya Oktaviani, Hollanda A Kusuma, Tianxiang Zhang, Jinya Su, Cunjia Liu, Wen Hua Chen, Hui Liu, Guohai Liu, M. Cavour, H. S. Duzgun, S. Kemec, D. C. Demirkan, Radhia Chairat, Yassine Ben Salem, Mohamed Aoun, Zolo Kiala, Onesimo Mutanga, John Odindi, and Kabir Peerbhay. Sentinel-2 data for land cover / use mapping: A review. *Remote Sensing*, 12:12291, 2020.
- 4 Hang Ren, Wei Guo, Zhenke Zhang, Leonard Musyoka Kisovi, and Priyanko Das. Population density and spatial patterns of informal settlements in nairobi, kenya. *Sustainability 2020, Vol. 12, Page 7717*, 12:7717, September 2020. doi:10.3390/SU12187717.
- 5 Yan Shi, Zhixin Qi, Xiaoping Liu, Ning Niu, and Hui Zhang. Urban land use and land cover classification using multisource remote sensing images and social media data. *Remote Sensing*, 11:2719, November 2019. doi:10.3390/RS11222719.
- 6 Aiman Soliman, Kiumars Soltani, Junjun Yin, Anand Padmanabhan, and Shaowen Wang. Social sensing of urban land use based on analysis of twitter users' mobility patterns. *PLOS ONE*, 12:e0181657, July 2017. doi:10.1371/JOURNAL.PONE.0181657.
- 7 Swapan Talukdar, Pankaj Singha, Susanta Mahato, Shahfahad, Swades Pal, Yuei An Liou, and Atiqur Rahman. Land-use land-cover classification by machine learning classifiers for satellite observations—a review. *Remote Sensing 2020, Vol. 12, Page 1135*, 12:1135, April 2020. doi:10.3390/RS12071135.
- 8 Linnet Taylor. What is data justice? the case for connecting digital rights and freedoms globally. *Big Data and Society*, 4, December 2017. doi:10.1177/2053951717736335.
- 9 Linnet Taylor and Dennis Broeders. In the name of development: Power, profit and the datafication of the global south. *Geoforum*, 64:229–237, August 2015.