Comparison of machine learning and statistical approaches for Digital Elevation Model (DEM) correction: interim results

Chukwuma Okolie[1,2,3*], Adedayo Adeleke[4], Julian Smit[5], Jon Mills[3], Iyke Maduako[6] and Caleb Ogbeta[7]

[1]Division of Geomatics, University of Cape Town, South Africa; chukwuma.okolie@alumni.uct.ac.za
[2]Department of Surveying & Geoinformatics, University of Lagos, Nigeria
[3]School of Engineering, Newcastle University, United Kingdom; jon.mills@newcastle.ac.uk
[4]Department of Geography, Geoinformatics and Meteorology, University of Pretoria, South Africa; adedayo.adeleke@up.ac.za
[5]Department of Civil Engineering and Geomatics, Cape Peninsula University of Technology, South Africa; smitj@cput.ac.za
[6]Department of Geoinformatics and Surveying, University of Nigeria, Nsukka, Nigeria; iykemadu84@gmail.com
[7]Geomatics Lab, School of Civil and Construction Engineering, Oregon State University, USA; ogbetac@oregonstate.edu

Commission II/WG4

Keywords: Digital elevation model, Copernicus GLO-30, ALOS World 3D, Multiple linear regression, Gradient boosted decision trees, Machine learning.

Abstract

The correction of digital elevation models (DEMs) can be achieved using a variety of techniques. Machine learning and statistical methods are broadly applicable to a variety of DEM correction case studies in different landscapes. However, a literature survey did not reveal any research that compared the effectiveness or performance of both methods. In this study, we comparatively evaluate three gradient boosted decision trees (XGBoost, LightGBM and CatBoost) and multiple linear regression for the correction of two publicly available global DEMs: Copernicus GLO-30 and ALOS World 3D (AW3D) in Cape Town, South Africa. The training datasets are comprised of eleven predictor variables including elevation, slope, aspect, surface roughness, topographic position index, terrain ruggedness index, terrain surface texture, vector ruggedness measure, percentage bare ground, urban footprints and percentage forest cover as an indicator of the overland forest distribution. The target variable (elevation error) was derived with respect to highly accurate airborne LiDAR. The results presented in this study represent urban/industrial and grassland/shrubland/dense bush landscapes. Although the accuracy of the original DEMs had been degraded by several anomalies, the corrections improved the vertical accuracy across vast areas of the landscape. In the urban/industrial and grassland/shrubland landscapes, the reduction in the root mean square error (RMSE) of the original AW3D DEM was greater than 70%, after correction. The corrections improved the accuracy of Copernicus DEM, e.g., > 44% RMSE reduction in the urban area and >32% RMSE reduction in the grassland/shrubland landscape. Generally, the gradient boosted decision trees outperformed multiple linear regression in most of the tests.

1. Introduction

Several methods have been proposed for correcting the elevation bias in digital elevation models (DEMs) for example, linear regression (Su & Guo, 2014; Olajubu et al., 2021; Pakoksung & Takagi, 2015; Preety et al., 2022). Other strategies have been proposed (e.g., Audenino et al., 2001; Bagheri et al., 2017, 2018; Bhardwaj et al., 2019; Fu et al., 2016; Okolie & Smit, 2022). Nowadays, supervised machine learning enables the modelling of complex relationships between variables, and has been deployed by researchers in a variety of fields (e.g., Hancock & Khoshgoftaar, 2020; Kotsiantis, 2007). Usually, the input layers into the models are the values of the elevation (derived from the DEM) and values of other DEM error-influencing parameters. While some researchers set the elevation error as the target layer (e.g., Kulp & Strauss, 2018; Liu et al., 2021), others utilized the groundtruth or reference DEM (RefDEM) as the target layer (e.g., Chen et al., 2020; Kasi et al., 2020; Kim et al., 2019).

In the existing literature, several studies have adopted either machine learning or statistical approaches in the task of DEM correction. However, to our knowledge, none of these studies have compared the performance of both approaches, especially with regard to publicly available global DEMs. Our previous work has already shown the potential of machine learning approaches, including gradient boosted decision trees (GBDTs) for DEM correction, e.g. (Okolie et al., 2023; Okolie et al., 2024) In this study, we share some results from the comparison of three recent implementations of gradient boosted decision trees (Extreme gradient boosting - XGBoost, Light gradient boosting - LightGBM and Categorical boosting - CatBoost), versus multiple linear regression (MLR) for enhancing the vertical accuracy of 30 m Copernicus and AW3D global DEMs in Cape Town, South Africa.

2. Methodology

2.1 Study area

Two different landscapes in Cape Town (South Africa; Figure 1) are presented in this assessment: urban/industrial (e.g. Figure 2a) and grassland/shrubland/dense bush (e.g., Figure 2b). The urban and industrial areas are located within the Cape Town metropolis, a large urban area with a high population density, industrial and business districts. Grassland/shrubland/dense bush includes natural or semi-natural grasses or low shrubs, open grassland, sparse bushland, transitional wooded grasslands and woodland areas, degraded vegetation with significantly reduced vegetation cover, natural or semi-natural areas dominated by trees and/or bushes, dense bush, closed woodland, thicket, scrub forest, dense shrubs and mangrove swamps (DFFE, 2023).

2.2 Digital elevation models

The 30 m Copernicus DEM (GLO-30) emanated from data acquired during the TanDEM-X Mission (Airbus, 2020). The Defense Gridded Elevation Data (DGED) format of the DEM was adopted. The 30 m ALOS World 3D (AW3D30) DEM is a derivative of the earlier high-resolution 5 m ALOS DEM (JAXA, 2023). AW3D version 3.2 was adopted in this research.

The reference DEM adopted is the 2 m City of Cape Town (CCT) airborne LiDAR-derived DEM acquired from the Information and Knowledge Management Department of the City of Cape Town (City Maps Office). The DEM is generated from LiDAR point clouds with a height accuracy of 15 cm. The data

*Corresponding author

acquisition for the dataset used in this study was conducted in phases between 2018 and 2021.

2.7 Model implementation and DEM correction

The input datasets are comprised of eleven predictor variables including elevation, slope, aspect, surface roughness, topographic position index, terrain ruggedness index, terrain

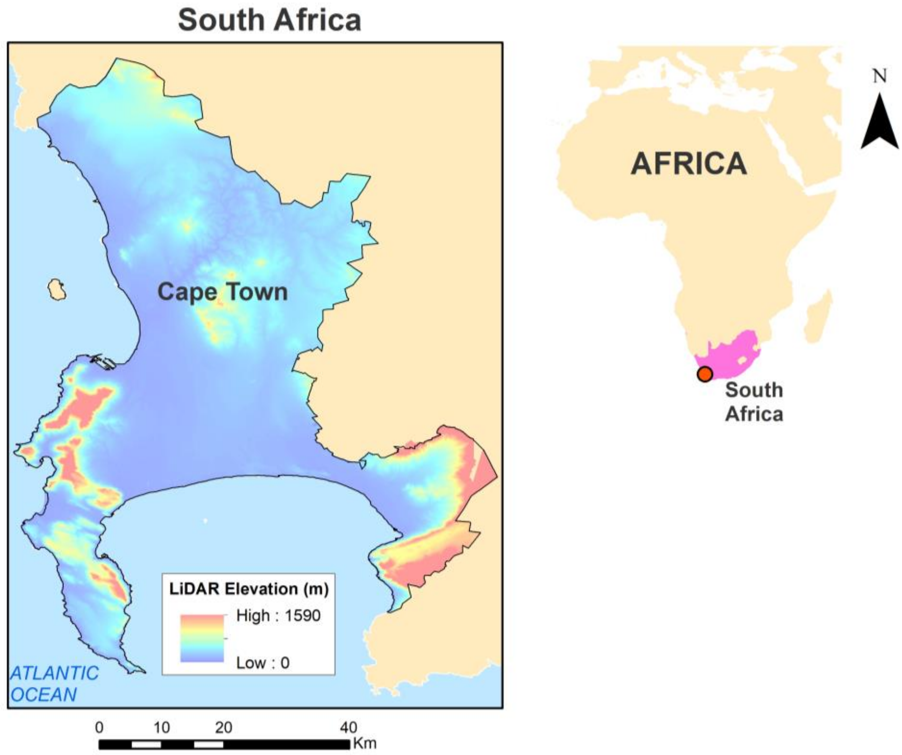

Figure 1. The location of Cape Town, South Africa

2.3 Global tree cover and bare ground

The global tree cover (treecover2010) and bare ground cover (bareground2010) provide estimates of the percent maximum tree canopy cover and the percent bare ground cover respectively (GLAD, 2023a; GLAD, 2023b; Hansen et al., 2013).

2.4 Global urban footprint

The Global Urban Footprint (GUF) is a database of human settlements that was introduced by the German Aerospace Centre (DLR) (Esch et al., 2012, 2018). The methodology for deriving the GUF is presented in Esch et al. (2010, 2013). The 0.4 arc-second (~12 m) 2012 GUF dataset was adopted in this research.

2.5 Terrain parameters

In this study, the following terrain parameters are considered: slope, aspect, surface roughness, topographic position index (TPI), terrain ruggedness index (TRI), terrain surface texture (TST), and vector ruggedness measure (VRM). The terrain parameters were generated within the QGIS 3.28.2 and SAGA GIS 7.8.2 software environments. For visualisation purposes, Figure 3 shows **s**lope and aspect maps of Cape Town derived from the airborne LiDAR-derived DEM.

2.6 Datum harmonisation

To achieve uniformity in the spatial reference systems, the global DEMs were projected from the Geographic to the Universal Transverse Mercator (UTM) coordinate system (UTM Zone 34 – southern hemisphere). The vertical datums of the DEMs were harmonised into the EGM2008 system.

surface texture, vector ruggedness measure, percentage bare ground, urban footprints and percentage forest cover. The target variable (elevation error) was derived with respect to highly

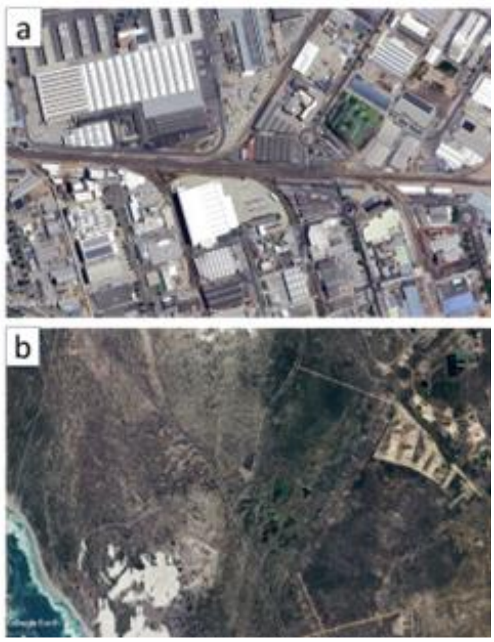

Figure 2. Satellite image views of some sections of the (a) urban/industrial, and (b) grassland/shrubland/dense bush landscapes

accurate airborne LiDAR. To check for multicollinearity in the input variables, Pearson's bivariate correlation analysis and the Variance inflation factor (VIF) were adopted. Pearson's bivariate correlation analysis was carried out to flag any significant correlations between the input variables. The VIF can indicate the problematic coefficients that are impacted by collinearity (Ferré, 2009). Using a general rule of thumb, variables with VIF > 10 are usually eliminated. In the case of multiple linear regression (MLR), surface roughness and TRI were flagged during multi-collinearity diagnostics and excluded from the input variables. Thus using MLR, the elevation error was expressed as a linear combination of nine input variables. Since multicollinearity is not a major concern for decision trees, all the eleven input variables were fed into the gradient boosted decision trees, GBDTs (XGBoost, LightGBM and CatBoost) where training was done using Python scripting in the Google Collaboratory environment. Generally, the models (trained with default hyperparameters) performed considerably well and demonstrated excellent predictive capability. Both models (GBDTs and MLR) were evaluated at several implementation sites for prediction and correction of DEM error. The corrections were achieved by subtracting the predicted elevation errors from the original elevations (i.e., $\mathrm{DEM}_{\mathrm{Corrected}} = \mathrm{DEM}_{\mathrm{Original}} - \Delta\mathrm{h}$).

3. Results and Discussion

In several instances after correction, the terrain offsets in the original DEMs were de-escalated (e.g. Figures 4 and 5). Table 1 compares the percentage reduction in RMSE of AW3D and Copernicus DEMs after correction. In the urban/industrial landscape, the RMSEs of the original AW3D DEM reduced by 72.1% (MLR), 72.2% (XGBoost), 72.8% (LightGBM) and 72.5% (CatBoost), while the RMSEs of the original Copernicus DEM reduced by 44.3% (MLR), 46.8% (XGBoost), 46.7% (LightGBM) and 47.0% (CatBoost). In the grassland/shrubland landscape, the RMSEs of the original AW3D DEM reduced by 72.6% (MLR), 72.6% (XGBoost), 73.2% (LightGBM) and 73.1% (CatBoost), while the RMSEs of the original Copernicus DEM reduced by 41.4% (MLR), 32.9% (XGBoost), 35.3% (LightGBM) and 32.3% (CatBoost).

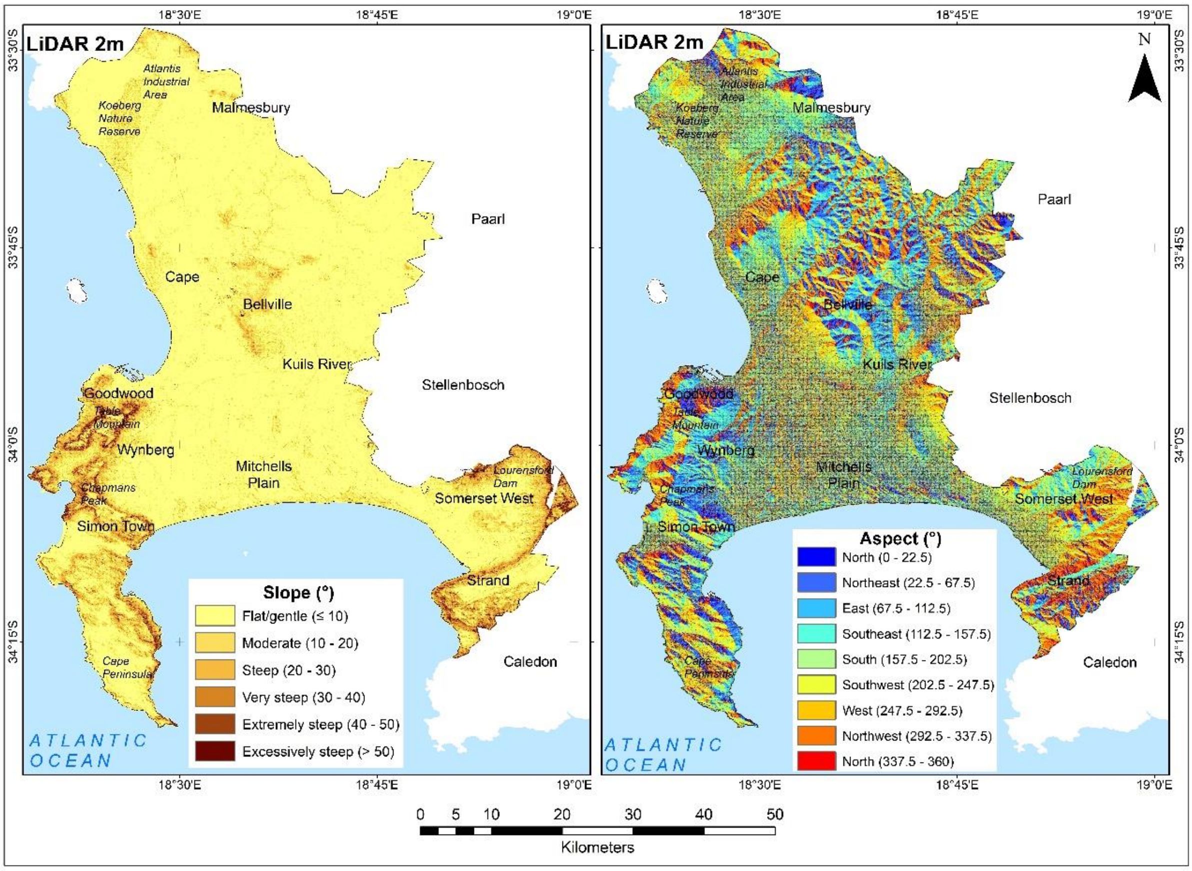

Figure 3. Slope (left) and aspect (right) classification derived from the airborne LiDAR-derived DEM

4. Conclusion

In the urban/industrial and grassland/shrubland landscapes, the reduction in RMSE of the original AW3D DEM was greater than 70%, after correction. The corrections improved the accuracy of Copernicus DEM, e.g., > 44% RMSE reduction in the urban area and >32% RMSE reduction in the grassland/shrubland landscape. The statistical-based (MLR) and machine learning (GBDT) correction achieved significant corrections of AW3D and Copernicus DEMs. While MLR outperformed the GBDTs in one scenario (i.e. Copernicus DEM in the grassland/shrubland landscape), the GBDTs outperformed MLR in most landscapes. The comparison proves the robustness of the GBDT-based correction in virtually all the landscapes under consideration. Future studies could integrate other approaches in the comparison.

Acknowledgements

The authors are grateful to the University of Cape Town and the Commonwealth Scholarship Commission for funding this research; and the Information and Knowledge Management Department, City of Cape Town for providing the LiDAR DEM. Also, we appreciate the assistance of Professor Jennifer Whittal, Dr Hossein Bagheri, and Tom Komar.

Landscape	% RMSE reduction (AW3D DEM)				% RMSE reduction (Copernicus DEM)			
	MLR	XGBoost	LightGBM	CatBoost	MLR	XGBoost	LightGBM	CatBoost
Urban/ industrial	72.1	72.2	72.8	72.5	44.3	46.8	46.7	47.0
Grassland/ shrubland	72.6	72.6	73.2	73.1	41.4	32.9	35.3	32.3

Table 1. Percentage reduction in RMSE of the original DEMs after correction

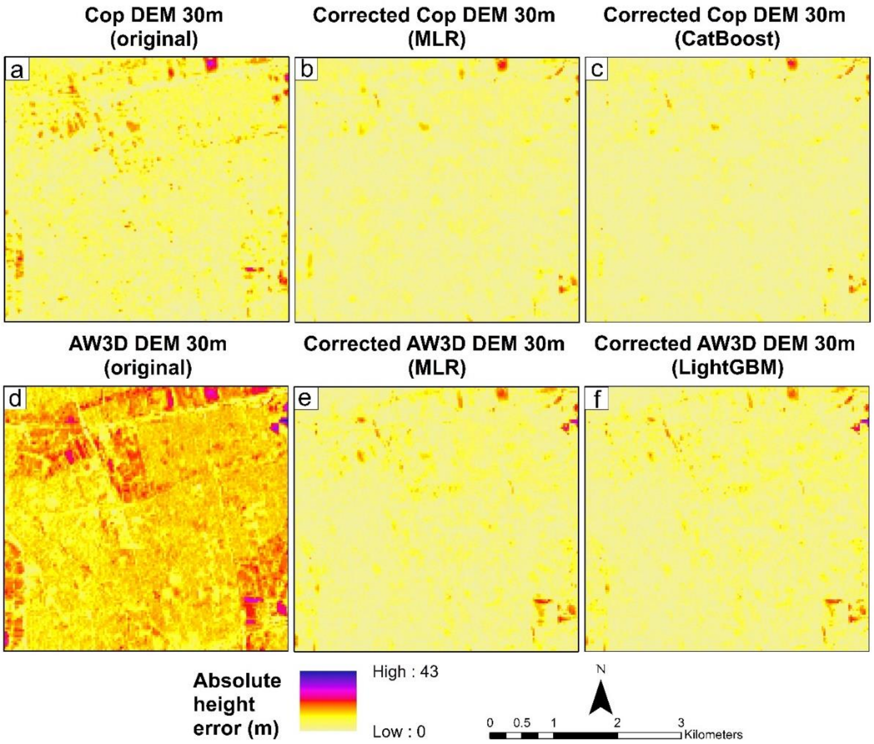

Figure 4. Absolute height error comparison of corrected DEMs in parts of the urban landscape

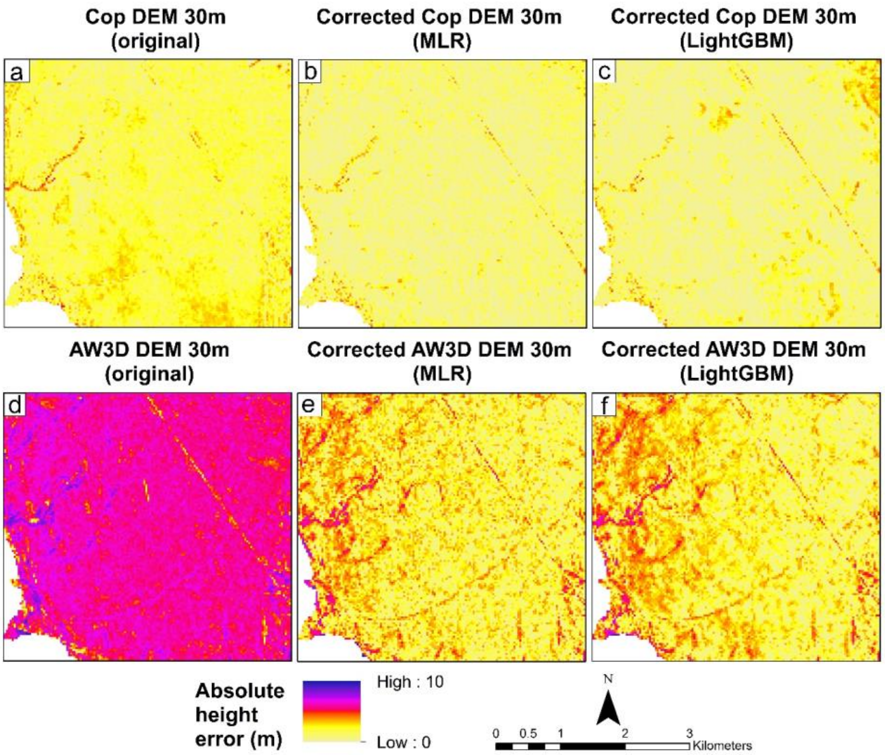

Figure 5. Absolute height error comparison of corrected DEMs in parts of the grassland/shrubland landscape

References

Airbus. 2020. *Copernicus DEM Copernicus Digital Elevation Model Product Handbook*.

Audenino, P., Rognant, L., Chassery, J. M., Planes, J. G. 2001. Fusion strategies for high resolution urban DEM. *IEEE/ISPRS Joint Workshop on Remote Sensing and Data Fusion over Urban Areas, DFUA 2001*, 90–94. https://doi.org/10.1109/DFUA.2001.985733

Bagheri, H., Schmitt, M., Zhu, X. 2017. Fusion of TanDEM-X and Cartosat-1 DEMS using TV-norm regularization and ANN-predicted weights. *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 3369–3372. https://doi.org/10.1109/IGARSS.2017.8127720

Bagheri, H., Schmitt, M., Zhu, X. X. 2018. Urban TanDEM-X Raw DEM Fusion Based ON TV-L1 and Huber Models. *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, 2018-July(June), 7251–7254. https://doi.org/10.1109/IGARSS.2018.8518870

Bhardwaj, A., Jain, K., Chatterjee, R. S. 2019. Generation of high-quality digital elevation models by assimilation of remote sensing-based DEMs. *Journal of Applied Remote Sensing*, 13(04), 44502. https://doi.org/10.1117/1.JRS.13.4.044502

Chen, C., Yang, S., Li, Y. 2020. Accuracy Assessment and Correction of SRTM DEM Using ICESat/GLAS Data under Data Coregistration. *Remote Sensing*, 12(20), 3435. https://doi.org/10.3390/RS12203435

DFFE. 2023. *SA National Land-Cover Datasets | Environmental Geographical Information Systems (E-GIS)*. https://egis.environment.gov.za/sa_national_land_cover_datasets

Esch, T., Bachofer, F., Heldens, W., Hirner, A., Marconcini, M., Palacios-Lopez, D., Roth, A., Üreyen, S., Zeidler, J., Dech, S., Gorelick, N. 2018. Where We Live—A Summary of the Achievements and Planned Evolution of the Global Urban Footprint. *Remote Sensing*, 10(6), 895. https://doi.org/10.3390/RS10060895

Esch, T., Marconcini, M., Felbier, A., Roth, A., Heldens, W., Huber, M., Schwinger, M., Taubenbock, H., Muller, A., Dech, S. 2013. Urban footprint processor-Fully automated processing chain generating settlement masks from global data of the TanDEM-X mission. *IEEE Geoscience and Remote Sensing Letters*, 10(6), 1617–1621. https://doi.org/10.1109/LGRS.2013.2272953

Esch, T., Taubenböck, H., Roth, A., Heldens, W., Felbier, A., Thiel, M., Schmidt, M., Müller, A., Dech, S. 2012. TanDEM-X mission—new perspectives for the inventory and monitoring of global settlement patterns. *Journal of Applied Remote Sensing*, 6(1), 61701–61702. https://doi.org/10.1117/1.JRS.6.061702

Esch, T., Thiel, M., Schenk, A., Roth, A., Müller, A., Dech, S. 2010. Delineation of Urban footprints from TerraSAR-X data by analyzing speckle characteristics and intensity information. *IEEE Transactions on Geoscience and Remote Sensing*, 48(2), 905–916. https://doi.org/10.1109/TGRS.2009.2037144

Ferré, J. 2009. Regression Diagnostics. *Comprehensive Chemometrics*, 3, 33–89. https://doi.org/10.1016/B978-044452701-1.00076-4

Fu, C. Y., Tsay, J.-R. 2016. Statistic tests aided multi-source DEM fusion. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 41(July), 227–233. https://doi.org/10.5194/isprsarchives-xli-b6-227-2016

Global 2010 Tree Cover (30 m) | GLAD, 2023. https://glad.umd.edu/dataset/global-2010-tree-cover-30-m

GLAD. 2023. *Global 2010 Bare Ground (30m)*. https://glad.umd.edu/dataset/global-2010-bare-ground-30-m

Hancock, J. T., Khoshgoftaar, T. M. 2020. CatBoost for big data: an interdisciplinary review. *Journal of Big Data*, 7(1), 1–45. https://doi.org/10.1186/S40537-020-00369-8/FIGURES/9

Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., Thau, D., Stehman, S. V., Goetz, S. J., Loveland, T. R., Kommareddy, A., Egorov, A., Chini, L., Justice, C. O., & Townshend, J. R. G. G. 2013. High-resolution global maps of 21st-century forest cover change. *Science*, 342(6160), 850–853. https://doi.org/10.1126/SCIENCE.1244693/SUPPL_FILE/HANSEN.SM.PDF

Dataset｜ALOS@EORC, 2023. https://www.eorc.jaxa.jp/ALOS/en/dataset/aw3d30/aw3d30_e.htm

Kasi, V., Yeditha, P. K., Rathinasamy, M., Pinninti, R., Landa, S. R., Sangamreddi, C., Agarwal, A., Dandu Radha, P. R. 2020. A novel method to improve vertical accuracy of CARTOSAT DEM using machine learning models. *Earth Science Informatics*, 13(4), 1139–1150. https://doi.org/10.1007/S12145-020-00494-1/TABLES/3

Kim, D.-E., Gourbesville, P., Liong, S.-Y. 2019. Overcoming data scarcity in flood hazard assessment using remote sensing and artificial neural network. *Smart Water 2019 4:1*, 4(1), 1–15. https://doi.org/10.1186/S40713-018-0014-5

Kotsiantis, S. B. 2007. Supervised Machine Learning: A Review of Classification Techniques. *Informatica*, 31(3), 249–268. https://www.informatica.si/index.php/informatica/article/view/148

Kulp, S., & Strauss, B. 2018. CoastalDEM: A global coastal digital elevation model improved from SRTM using a neural network. *Remote Sensing of Environment*, 206, 231–239. https://doi.org/10.1016/J.RSE.2017.12.026

Liu, Y., Bates, P. D., Neal, J. C., Yamazaki, D. 2021. Bare-Earth DEM Generation in Urban Areas for Flood Inundation Simulation Using Global Digital Elevation Models. *Water Resources Research*, 57(4), e2020WR028516. https://doi.org/10.1029/2020WR028516

Okolie, C., Adeleke, A., Mills, J., Smit, J., Maduako, I., Bagheri, H., Komar, T., & Shidong Wang. 2024. Assessment of explainable tree-based ensemble algorithms for the enhancement of Copernicus digital elevation model in agricultural lands. *International Journal of Image and Data Fusion*, 1–31.

https://doi.org/10.1080/19479832.2024.2329563

Okolie, C. J., Smit, J. L. 2022. A systematic review and meta-analysis of Digital elevation model (DEM) fusion: pre-processing, methods and applications. *ISPRS Journal of Photogrammetry and Remote Sensing*, 188, 1–29. https://doi.org/10.1016/J.ISPRSJPRS.2022.03.016

Okolie, C., Mills, J., Adeleke, A., Smit, J. 2023. *Digital elevation model correction in urban areas using extreme gradient boosting, land cover and terrain parameters*. https://arxiv.org/abs/2308.06545v1

Okolie, C., Mills, J., Adeleke, A., Smit, J., & Maduako, I. 2023. The explainability of gradient-boosted decision trees for digital elevation model (DEM) error prediction. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-M-3–2023(M-3–2023), 161–168. https://doi.org/10.5194/ISPRS-ARCHIVES-XLVIII-M-3-2023-161-2023

Olajubu, V., Trigg, M. A., Berretta, C., Sleigh, A., Chini, M., Matgen, P., Mojere, S., Mulligan, J. 2021. Urban correction of global DEMs using building density for Nairobi, Kenya. *Earth Science Informatics*, 14(3), 1383–1398. https://doi.org/10.1007/S12145-021-00647-W/TABLES/5

Pakoksung, K., Takagi, M. 2015. Digital elevation models on accuracy validation and bias correction in vertical. *Modeling Earth Systems and Environment*, 2, 1–13. https://doi.org/10.1007/s40808-015-0069-3

Preety, K., Prasad, A. K., Varma, A. K., El-Askary, H. 2022. Accuracy Assessment, Comparative Performance, and Enhancement of Public Domain Digital Elevation Models (ASTER 30 m, SRTM 30 m, CARTOSAT 30 m, SRTM 90 m, MERIT 90 m, and TanDEM-X 90 m) Using DGPS. *Remote Sensing*, 14(6), 1334. https://doi.org/10.3390/RS14061334/S1

Su, Y., Guo, Q. 2014. *A practical method for SRTM DEM correction over vegetated mountain areas*. 87, 216–228. https://doi.org/10.1016/J.ISPRSJPRS.2013.11.009