reviewed paper

Spatial Determinants of Urbanisation in Debre Markos, Ethiopia: Modelling Building Footprint

Moges Wubet Shita, Haile Legese Zewale, Gerhard Navratil

(Moges Wubet Shita, MA, Department of Geodesy and Geoinformation, TU Wien, Vienna, Austria, and Institure of Land Administration, Debre Markos University, Ethiopia, moges.shita@geo.tuwien.ac.at)
(Haile Legese Zewale, MSc, Department of Geodesy and Geoinformation, TU Wien, Vienna, Austria, and Institure of Land Administration, Debre Markos University, Ethiopia, haylea.zewale@geo.tuwien.ac.at)
(Dipl.-Ing. Dr. Gerhard Navratil, Department of Geodesy and Geoinformation, TU Wien, Vienna, Austria, gerhard.navratil@geo.tuwien.ac.at)

1 ABSTRACT

Urbanization puts pressure on socio-economic and environmental aspects worldwide. Spatial factors play a pivotal role in driving this urbanization; for instance, geographical features, proximity to socio-economic services, and government zoning regulations are spatial determinants discussed in the literature. The purpose of this study is to understand the spatial determinants of urbanization based on building footprints.

The Google Open Building Dataset has been used for retrieving building footprints. Additionally, 26 dependent variables were collected from various sources. Road networks were extracted from OSMnx, and geographical data were collected from Google Earth Engine. Points of interest for proximity estimation were gathered from the Debre Markos municipality, and some socio-economic data were collected through a survey of 385 respondents, which were then interpolated to the entire area. The independent variables are categorized as geographical, proximity, socio-economic, and governmental regulation factors. About 25,000 training samples were extracted from each variable to train the models.

Two methods were employed in this research: the binary logistic regression and the machine-learning model of XGBoost. Binary logistic regression was employed for its interpretability, while XGBoost was employed for its superior data management and prediction accuracy. According to the results of the area under the curve (AUC) for accuracy measurement, logistic regression achieved 0.73, and XGBoost achieved 0.82. However, the data fit the model in both cases. Distance from road, building height zone, road network density, and slope are among the top factors determining urban building footprint. This implies that the likelihood of building has increased near roads. The results of building-height zoning show that local government regulations affect the likelihood of building, and a model result on slope also indicates that topography is a significant determinant of urbanization.

Keywords: XGBoost, logistic regression, urban expansion, building presence, urban modeling

2 INTRODUCTION

The primary drivers of urbanization are natural population growth and rural-to-urban migration (Arif & Gill, 2023; Ayele & Tarekegn, 2020). Location, development level, country size (Arif & Gill, 2023), topographic, planning and policy constrients (Wu et al., 2015), and economic level determine the level of urban expansion (Dutta et al., 2020). Urbanization is a global concern, exerting pressure on social, economic, and environmental issues in both the global North and the global South (Arif & Gill, 2023; Braimoh & Onishi, 2007). The fast-growing and uncontrolled expansion causes nuanced problems (Angel, 2023), such as environmental degradation, unplanned urban housing and lack of infrastructure (Karimi et al., 2021). It has also consumed agricultural land, leading to food insecurity, soil erosion, social injustice, and high housing burden (Chen et al., 2016).

Although cities can serve their residents through densification or horizontal expansion, the focus in recent years has been predominantly on horizontal urban expansion (Angel, 2021). In Africa, urban centers have expanded into peri-urban areas, where infrastructure is limited and social services are inaccessible. The residents become poor, and the areas become sources of various environmental pollutants (Braimoh & Onishi, 2007). Ethiopia experienced such a rapid urbanization. Inadequate planning or problems in its implementation generally characterize this kind of urbanization. The result of such a velopment are cities lacking the necessary infrastructure and services (Ayele & Tarekegn, 2020).

Spatial determinants of urbanization are driving factors that emphasize the spatial dependency between geographic features (Chen et al., 2016). The spatial determinants of urbanization used by previous studies are:

- Social infrastructure accessibility (Community space, place of worship, social business, and parks) (Fraser et al., 2024),
- Proximity of road, land value, proximity of an education center, proximity of a utility service center, elevation, proximity of the old city, distance to the canal, elevation, distance from rivers, and distance from a forest (Sarkar & Chouhan, 2020).
- Transportation networks and proximity to the city center (Sarkar & Chouhan, 2020),
- Topological factors, neighborhood factors, socio-economic factors, and spatial planning policies (Mustafa, et al., 2018)
- Accessibility and topographic features (Christensen & McCord, 2016),
- Market access and terrain slope (Christensen & McCord, 2016),
- Proximity variables, topological variables, density, banks, company offices, stores, and transportation services (Chen et al., 2016),

In Debre Markos, rapid urbanization and spatial expansion are transforming non-urban land uses into urban uses. This expansion is driven by natural population growth and by rural-to-urban migration for expectation of better life and work opportunity, which increases demand for housing and infrastructure. Ultimately, this leads to challenging urban planning and development.

Policy interventions are essential to mitigate the destructive impact of built-up developments on the environment, natural resources, and human health (Mustafa, Heppenstall, et al., 2018). Thus, a good understanding of the factors of urbanization is vital for developing strategies for various purposes, such as development planning (Hofmann & Wan, 2013). This study aims to understand the spatial factor of urbanization and this knowledge will contribute to the city's spatial planning.

Although there are several studies on urban growth in Debre Markos, most rely on land-use land-cover classification and boundary expansion (Agegnehu et al., 2015; Abebaw, 2017; Temesgen & Huseyin, 2019). However, the spatial determinants of urban growth, including proximity factors, government regulations, socio-economic, and geographical factors, remain insufficiently studied. Analyzing these factors and their implications can significantly enhance urban planning and development in the city.

Besides, the authors recognized a shortage of studies that utilize building footprint data and employ both XGBoost spatial machine learning and econometric binary logistic models. The research is based on Angel's proposed future research ideas (2023). Among the many suggested questions, the authors were particularly motivated by two key questions: What are the determinants of urban footprints in cities? And what are the drivers of urban expansion? Based on these research questions, the following research questions were developed, considering the availability of data:

- What physical, socio-economic, and policy factors determine the likelihood of building footprint?
- How are the spatial factors correlated with each other?

3 RESEARCH METHODS AND MATERIALS

3.1 Description of the study area

Debre Markos, illustrated in Figure 1, the capital of the East Gojjam Administrative Zone, is located 300 kilometers northwest of Ethiopia's capital, Addis Ababa, and 265 kilometers from Bahir Dar in the Amhara National Regional State. The city's population has grown significantly in the last years due to rural-to-urban migration and natural increase. It was about 92,500 in 2014, and is estimated to be about 161,000 in 2025 (https://ess.gov.et/).

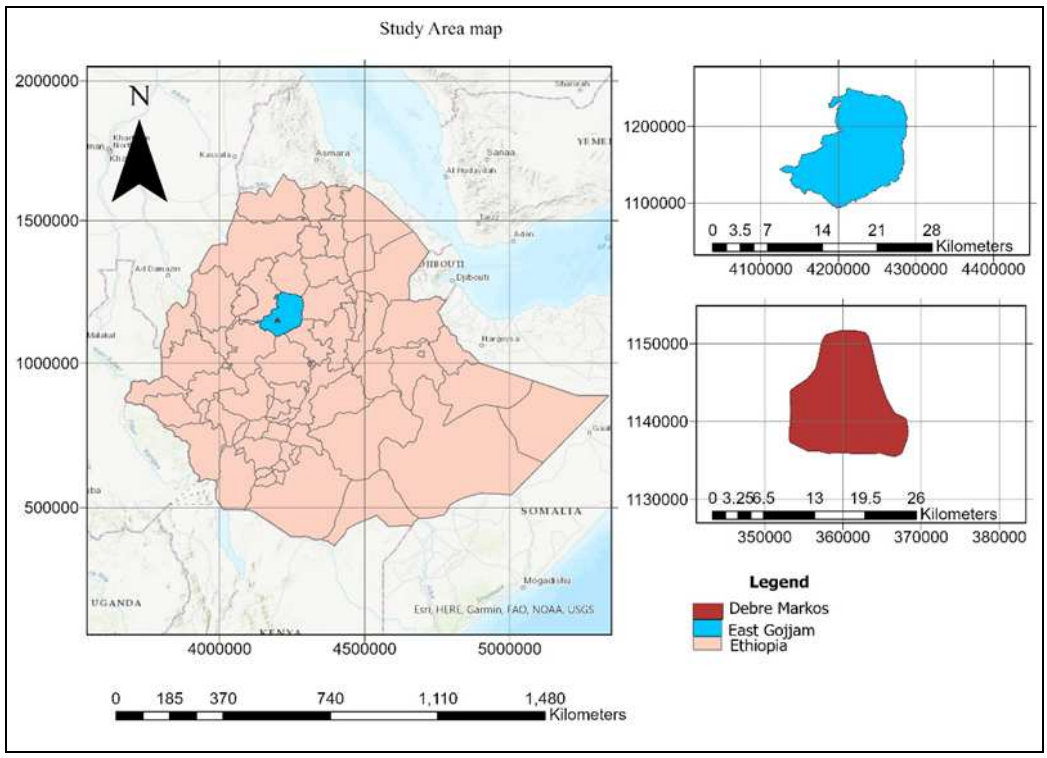

Figure 1: Study area map; source: "Rental housing affordability and spatial variation in rapidly growing urban areas of Ethiopia: evidence from Debre Markos city" submitted to the "International Journal of Housing Markets and Analysis (IJHMA)" by Moges Wubet Shita et al.

3.2 Data Source

The source for building footprint presence is open buildigs. It is a temporal annual dataset with an effective resolution of 4 meters, detailing building presence, fractional building counts, and building heights. The data are used in the fields of urbanization, environmental science, population mapping, public health planning, humanitarian, and disaster response (https://sites.research.google/gr/open-buildings/). The region-of-interest data for the building footprint can be downloaded from Google Earth Engine. In developing nations, alternative data sources are often based on remotely sensed data (see, for example, Sirko et al., 2021). The Figure 2 visualizes the difference in building footprint presence around the center of Debre Markos and it's perphery.

The independent variables are classified in four categories (Table 1): Government regulatory variables, socio economic variables, geographical variables, and proximity variables.

Land value grade and building height zoning are government regulatory variables selected for this study. The source of the variables is Debre Markos municipality (2024). Land value grades were provided in the DWG (AutoCAD drawing file) format and then converted into raster format for the simplicity of extracting the pixel value. Building height zone data is provided in raster format already. Land value, rental price per unit, household income, household affordability, and household size are socio-economic data. The land value per square meter is determied by interpolating from the price of 769 parcels from the Debre Markos municipality, wich were adjusted to make it in the same year. After adjustment, Inverse Distance Weighted (IDW) interpolation was used to obtain a continuous value field for the entire area. The other socio-economic variables were obtained by a survey of 385 households.

The geographical variables, elevation, slope, aspect, and hill shade are obtained from an image collection of CGIAR/STRM90_V4 throgh Google Earth Engine (GEE). Initially, the resolution was 90 meters; it was then resampled to 10 meters, to obtain a similar cell size to the other factors. To estimate proximity to the road network and road density, the road network was extracted from OSMnx, and then the proximity to the road network and road density were computed. The drainage line network, primary schools, secondary schools, streetlights, police stations, market centers, health centers, playing grounds, and green areas were collected from the Debre Markos municipality. Areas of worship and the city center pointwere extracted from Google Earth maps. Finally, the Euclidean distance of each pixels to each point of interest was computed.

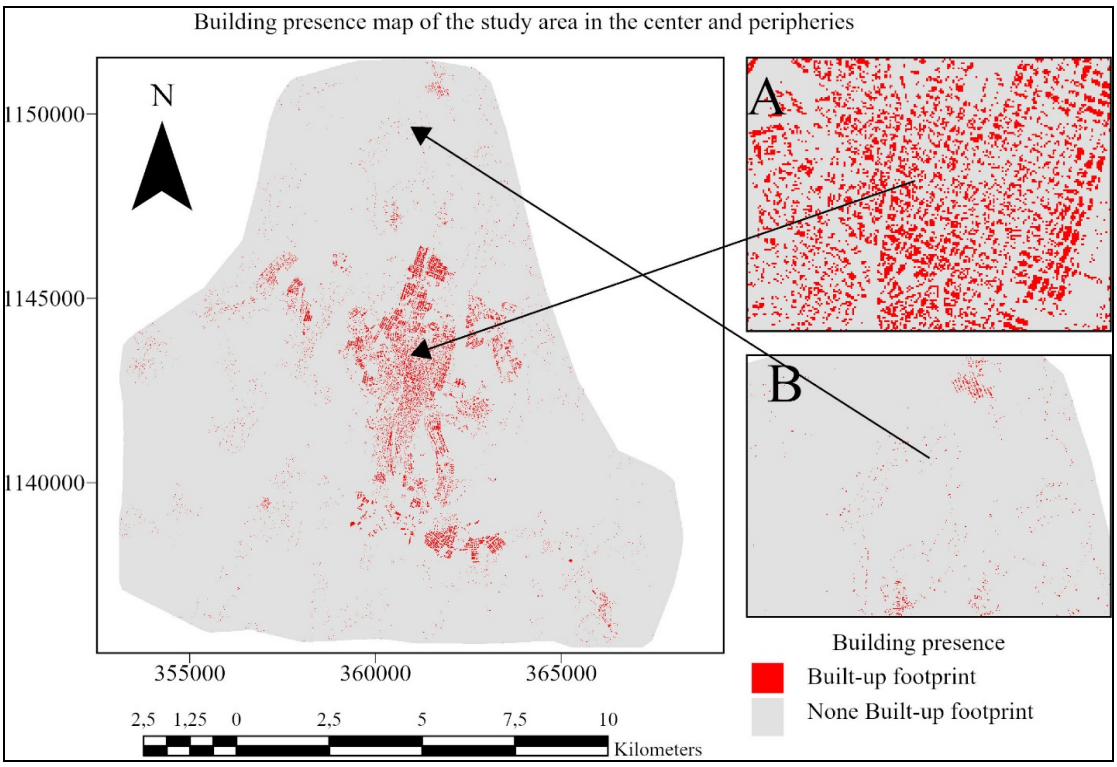

Figure 2: Building footprint presence of Debre Markos extracted open buildings collections using GEE.

Finally, about 25,000 random points were generated within the Debre Markos city map. The values of all 26 variables are extracted for these poins to create training data.

Variable Category	Representation	Variable name	Type of data	Source
Dependent Variable	BP	Building presence	Binary	Open buildings
Government regulatory variables	LVG	Land value grade	Discrete	DM municipality
	BHZ	Building height zoning	Discrete	DM municipality
Socio-economic variables	LV	Land value	Contentious	DM municipality*
	RPPU	Rental price per unit	Contentious	From 385 survey data*
	HHI	Household income	Contentious	From 385 survey data*
	AHH	Household Affordability	Contentious	From 385 survey data*
	HHS	Household size	Contentious	From 385 survey data*
Geographical variables	Slope	Slope	Contentious	DEM from GEE
	Elevation	Elevation	Contentious	DEM from GEE
	Aspect	Aspect	Contentious	DEM from GEE
	Hill shade	Hill shade	Contentious	DEM from GEE
Proximity and accessibility variables	DWA	Distance from worship areas	Contentious	From Google Earth**
	DfD	Distance from drainage line	Contentious	DM municipality**
	DPS	Distance from primary school	Contentious	DM municipality**
	DSS	Distance from secondary school	Contentious	DM municipality**
	DSL	Distance from street lite	Contentious	DM municipality**
	DRN	Distance from road network	Contentious	OSMnx**
	DPLS	Distance from police stations	Contentious	DM municipality**
	DMC	Distance from Market Center	Contentious	DM municipality**
	DHC	Distance from the Health Center	Contentious	DM municipality**
	DfG	Distance from green spaces	Contentious	DM municipality**
	DPG	Distance from playing ground	Contentious	DM municipality**
	DCC	Distance from the city center	Contentious	Google Earth **
	WLD	Water line density	Contentious	DM municipality ***
	RND	Road network density	Contentious	OSMnx***
	DD	Drainage line density	Contentious	DM municipality ***
Note: After collecting the raw data, we preprocessed it. *Indicates the variable in which the values were interpolated to have a continuous value in the map area. ** indicates that the Euclidean distance has been computed to the point of interest. *** indicates line density was calculated. All data were resampled to a raster width of 10x10 meters.				

Table 1: Variables selected for this study based on literature review and classified in four categories.

3.3 Model specification

This research aims to identify the determinants of urbanization by modeling the presence of building footprints. Since our data indicate that the presence of a building footprint is binary, logistic regression was adopted for the econometric model and extreme gradient boosting (XGBoost) for machine learning. Logistic regression is a generalized linear model, based on the assumption of linearity. It models the probability of

observing binary responses (0 or 1) for the response variable (Starbuck, 2023; Shahri et al., 2021). XGBoost performs well even with unbalanced data. Logistic regression on the other hand performs poorly on unbalanced datasets (Shahri et al., 2021). However, interpreting the logistic regression model results is easier that interpreting the XGBoost model (Pesantez-Narvaez et al., 2019).

Hyper parameter	Search space	Description
No. of estimators	200, 300, 500, 800, 750, 1000, 1500, 2000	Number of boosting rounds
learning rate	0.001, 0.01, 0.025, 0.05, 0.075, 0.2, 0.3, 0.4, 0.5, 0.6	Learning rate of model
Max depth	9, 10, 11, 12, 13, 14, 15, 16	Maximum depth of the trees
Min child weight	4, 5, 6, 7, 8, 9	Minimum sum of instance weight in a child
Gamma	0.1, 0.2, 0.3, 0.4	Minimum loss reduction for further partitioning
Subsample	0.5, 0.6, 0.7, 0.8, 0.9	Ratio of training data used per tree
Colsample bytree	0.6, 0.7, 0.8, 0.9, 1.0	Ratio of columns subsampled per tree
Objective	binary: logistic	Objective function for optimization
Booster	gbtree, gblinear	Boosting method used
Eval metric	Auc	Area under the curve from the AOC curve for data validation

Table 2: Search space for hyper parameter for XGBoost Model (source: Kmen et al., (2024)

Both models were implemented in Python. In binary logistic regression, data were transformed using the standard scalar method. Then, the multicollinearity test was performed using the Variance Inflation Factor (VIF). Based on the VIF results, several variables are highly correlated. The VIF values above 10 are dropped from the model (Cheng et al., 2022). For XGBoost training, the data was split into test and training sets at 30% and 70%, respectively. A 10-fold validation with a randomized search was employed, with 500 iterations used to tune the hyperparameters.

After training of the model the best parameter were estimated as “subsample”: 0.6, “objective”: “binary: logistic”, “n_estimators”: 2000, “min_child_weight”: 8, “max_depth”: 11, “‘learning_rate”: 0.01, “gamma”: 0.3, “colsample_bytree”: 0.7, “booster”: “gbtree”.

3.4 Model Performance Evaluation

There are several performance measurements, such as R2, mean squared error, and Receiver Operating Characteristic (ROC) curve. However, its usage depends on the nature of the model. Since our model is a binary prediction, the ROC curve and AUC are selected to evaluate the model's performance, as Nakas et al. (2023) elaborated. Ideally, the ROC curve should touch the top-left corner, increasing the area under the curve and indicating better accuracy (Nakas et al., 2023). ROC curves have been used to compare different models (James et al., 2023). ROC is an outstanding method for estimating a model's quality by comparing a binary map (actual) with a probability map for the likelihood of change (Mustafa et al., 2018).

The AUC is a direct and convenient overall measure to evaluate the model’s performance (Zou et al., 2012). Eventually, it is summarized into a single number (Nathalie Japkowicz & Mohak Shah, 2011). If AUC is less than or equal to 0.5, the performance is no better than chance (James et al., 2023). If it is 1, it indicates perfect classification (Mustafa et al., 2018). Therefore, the value of RUC should lie between 0.5 and 1, howver a value approches to 1 have best performance, indicating that the model is better than random discrimination.

According to Nathalie Japkowicz & Mohak Shah (2011), RUC is defined as,

$$RUC = \frac{\sum_{i=1}^{|Tp|}(Ri-i)}{|Tp||Tn|} \quad (1)$$

where Tp ⊂ T and Tn ⊂ T are, respectively, the subsets of positive and negative examples in test set T, and Ri is the rank of the i^{th} example in Tp given by classifier f.

4 RESULT

Figure 3 shows the correlational analysis of the spatial variables. Red numbers indicate a strong relationship, blue numbers a weak relationship. Obviously, the proximity variables are strongly correlated. The proximity distances increase or decrease simultaneously. This is because infrastructure and social services are concentrated around specific locations. For instance, health centers, water lines, market centers, schools, and worship areas are built near roads, street lights, playing grounds, and the city center. Green spaces are established near roads in the city center.

Figure 3 also depicts the correlation between building presence (BP) and the predictor variables. Building height zoning (BHZ), land value (LV), rental price per unit (RPPU), household affordability (AHH), elevation, water line density (WLD), road network density (RND), and drainage density (DD) are positively correlated. The remaining variables, for example, land value grade (LVG), household income (HHI), household size (HHS), slope, distance from market center (DMC), distance from health center (DHC), distance from playing ground (DPG), and distance from the center of the city (DCC) are negatively correlated.

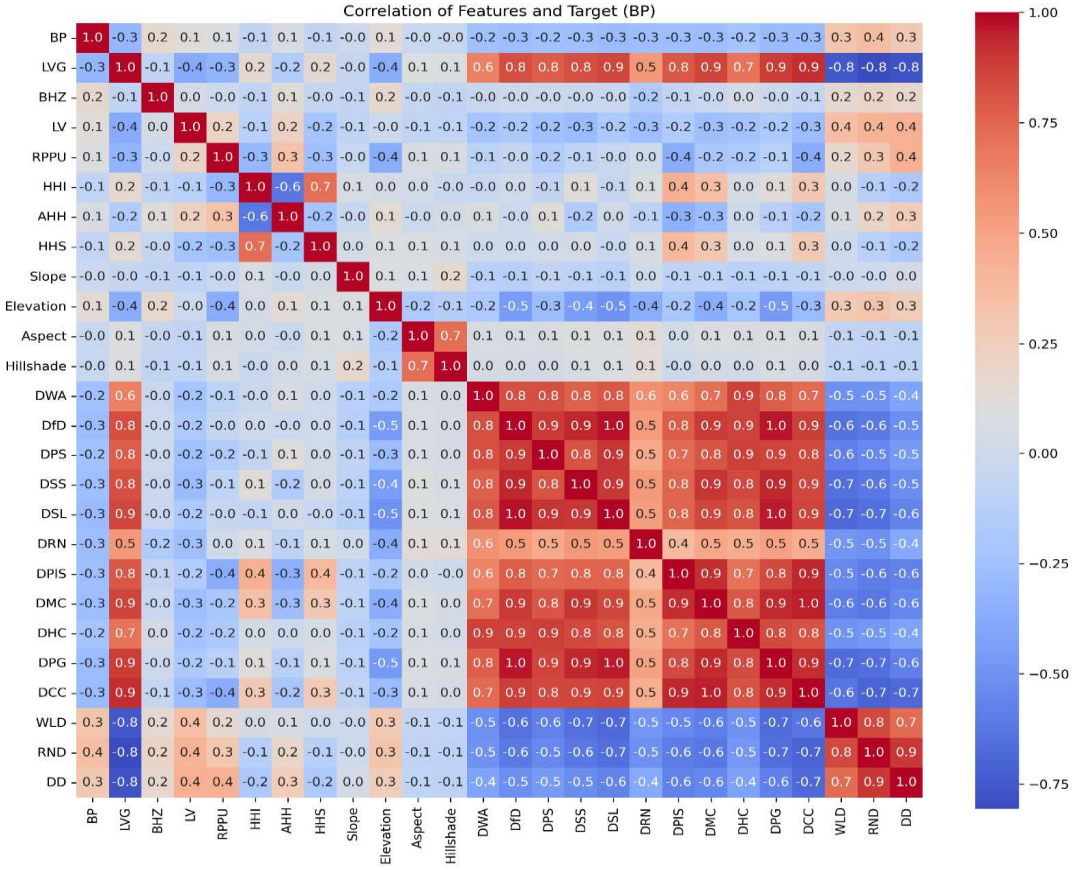

Figure 3: Correlation of variables.

As illustrated in Figure 3 of the ROC curve, the performance measures for the two models are: XGBoost scores 0.82 and logistic regression scores 0.73. In both models, AUC exceeds 0.5, which means the models perform better than random guessing. However, XGBoost performs better than a binary logistic regression in this case.

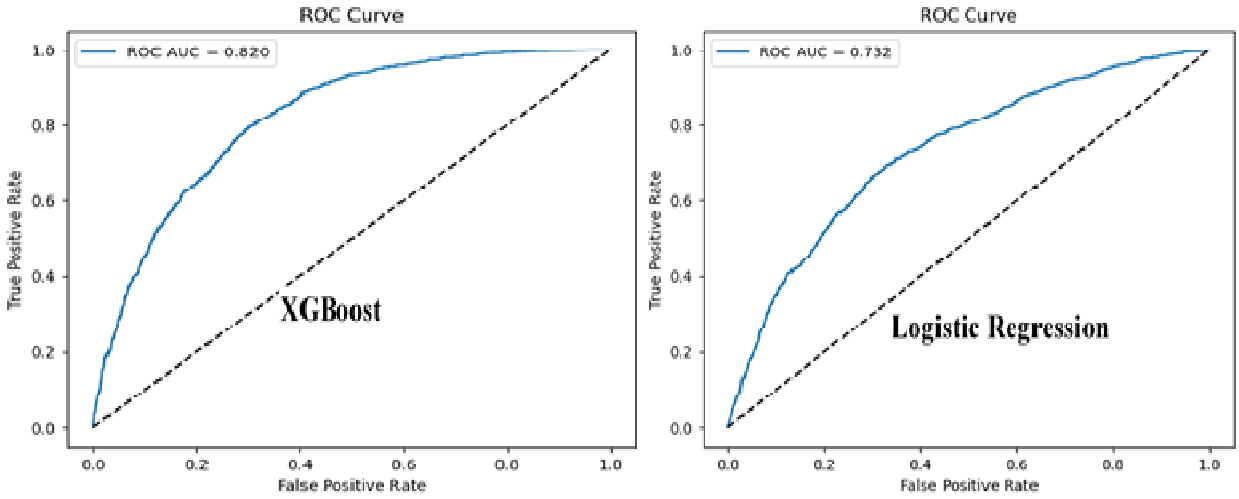

Figure 4: ROC curve of the two models.

Feature	Coefficient	Odds Ratio	Marginal Effect	p-value
Constant	-1.567	0.209		0.000*
DRN	-0.572	0.564	-0.046	0.000*
DPlS	-0.321	0.725	-0.012	0.000*
BHZ	0.284	1.328	0.000	0.000*
WLD	0.234	1.263	0.014	0.000*
DSS	0.149	1.161	0.082	0.001*
AHH	-0.109	0.897	-0.010	0.000*
HHI	-0.087	0.917	-0.016	0.005*
DWA	-0.086	0.917	-0.004	0.039**
Slope	-0.085	0.918	-0.010	0.000*
DHC	-0.083	0.920	-0.033	0.082***
Elevation	0.071	1.073	0.001	0.016**
HHS	0.071	1.073	0.012	0.009*
DPS	0.029	1.029	0.021	0.577
Hill shade	0.013	1.013	0.012	0.643
Aspect	0.004	1.004	0.002	0.888
LV	-0.002	0.998	-0.000	0.914
RPPU	-0.002	0.998	-0.012	0.921

Table 3: Logistic regression result, which shows the significance of variables (*represents variables that are significant at p <0.01, and ** represents variables that are significant on p<0.05, and *** represents variables that are significant at p<0.1)

Nine of the 26 variables were dropped from the model due to multicollinearity. Table 3, shows the results of the logistic regression. DPlS, BHZ, WLD, DSS, AHH, slope, and HHS are significant variables at (p<0.01), DWA and elevation are significant at (p<0.05), and DHC is substantial at (p<0.1). However, DPS, hill shade, aspect, LV, and RPPU are not significant determinants of building footprint presence. An increase of DRN, DPlS, AHH, HHI, DWA, Slope, and DHC affect the likelihood of building presence by reducing the probability of a building footprint. However, increasing values in BHZ, WLD, DSS, Elevation, and HHS raise the likelihood of a building footprint.

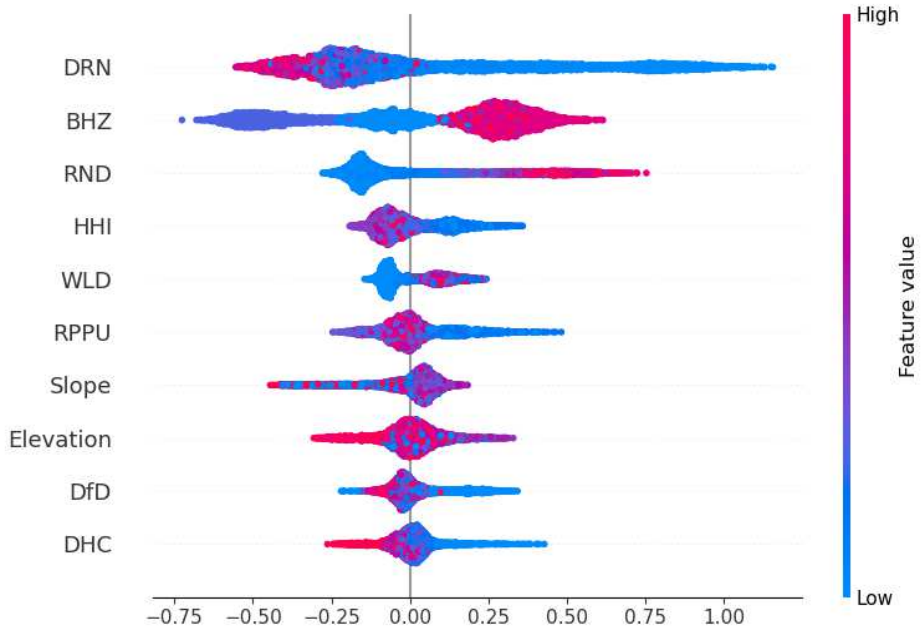

Figure 5: Top ten Shapley values of a feature that shows the impact of the variable on the model output from XGBOOST

”Shapley Additive exPlanations (SHAP) adds global and local interpretability to ML-based models by computing the marginal contribution of each feature” (Movsessian et al., 2022). This SHAP value is used to understand the feature contribution to the result. According to the SHAP values of XGBoost (see Figure 5), distance from road (DRN), building height zoning (BHZ), road network density (RND), household income (HHI), water line density (WLD), rental per unit price (RPPU), slope, elevation, distance from green spaces (DfG), and distance from health center (DHC) are the ten variables that influence the likelihood of building presence most. DD, LVG, DCC, DMC, AHH, DWA, DSS, hill shade, DPG, and DPS are contributing the least to the model output.

5 DISCUSSION AND IMPLICATIONS

The purpose of this study is to identify the spatial determinants of urbanization using building footprints. To meet the objective of this study, econometric logistic regression and the machine-learning model XGBoost were employed. The results reveal that XGBoost outperforms binary logistic regression.

Off-government regulatory variable building height zoning (BHZ) is one of the significant variables in the binary logistic regression, with a positive association with the likelihood of building presence. An increase in building height increases the likelihood of building availability. The XGBoost result also confirms that the BHZ is the second most contributing factor for determining the building presence. From this, we can see how local government regulations affect urbanization, shaping and expanding cities. The effect of policy on urbanization is also documented (Braimoh & Onishi, 2007).

Household income (HHI), household housing affordability (AHH), and household size (HHS) are significant socio-economic variables in a binary logistic regression model. Household income is inversely correlated with the presence of a building footprint and is the fourth contributing factor in the XGBoost model. As income increases, the household needs more space and can afford the commuting costs to the city center. Thus, they can search spaces outside the city center. Housing affordability is a significant factor affecting the probability of a building footprint, with the likelihood decreasing as household affordability increases. However, there are high commuting costs. For many, houses are only affordable outside the city center, where buildings are relatively compact. Household size has a positive coefficient: an increase in household size increases the likelihood of building presence. From these, it can be inferred that socio-economic factors are determinants of urbanization in shaping the presence of buildings, whether compact or sparse.

Among the topographical variables, elevation and slope are significant factors that affect the likelihood of building in urban areas. According to this study, these two variables are among the top contributors to the XGBoost model. Similarly, in the binary logistic regression model, elevation is positively associated with the presence of a building footprint. A slope negatively affects the likelihood of building and urbanization is expanding at the plateau of the city. As the slope increases, it exposes the site to runoff, increases maintenance and construction costs, and makes infrastructure expansion difficult. Therefore, topography has an effect on urbanization by shaping the urban expansion.

Among the proximity variables, DRN, DPlS, AHH, HHI, DWA, and DHC affect the likelihood of building presence; a one-unit increase lowers the probability of a building footprint. However, WLD and DSS, have increased the likelihood of a building footprint. In addition, the results from XGBoost indicate that DRN and RND are the most influential variables for the likelihood of building footprint.

Urbanization depends on the expansion of road infrastructure and the distribution of social services, such as markets, health centers, and places of worship. The effect of distance to roads on urbanization and urban expansion has been documented in several studies (e.g., Mustafa, Heppenstall, et al., 2018; Chen et al., 2016; Sarkar & Chouhan, 2020). The effects of policy and proximity variables are documented in Braimoh & Onishi (2007). Furthermore, the topography of the land, such as slope and elevation, and socio economic facors has also determined the urbanization that shapes the pattern of buildings and the city's directional growth. As it has also been documented by Jing et al. (2022). The effect of road density on urban expansion was explained by Hofmann & Wan (2013). Therefore, urbanization is determined by the interplay of government policies and regulations, geographical features, accessibility to infrastructure and social services, and socio-economic factors.

6 CONCLUSION

In this study, correlational analysis of the variables was employed. Several variables were found to be strongly correlated with each other. This analysis extended to examine the relationship between building presence and 26 independent variables. Binary logistic regression was employed to assess the statistical significance of the independent variable and to facilitate a more straightforward interpretation. The XGBoost machine learning method was used for its predictive power and to draw the contribution of the variables to the model using SHAP values. According to the model's performance metrics, XGBoost performs better than logistic regression. Of the independent variables, distance to road, building zoning, and road network density are the most determinate factors of urbanization.

This study has the following limitations:

- Due to the availability of open building data, this study is cross-sectional. However, it would be more generalizable if it included different periods.
- The socio-economic data are interpolated from the survey of 385 respondents; the result would be good if the data is based on the detailed survey of the population
- The land value data is interpolated from the hundreds of parcel values

Funding: Austrian partnership program in higher education and research for development (APPEAR) and a program of Austrian development cooperation (ADC) and implemented by Austria's Agency for Education and Internationalization (OeAD GmbH) (0894-01/2020).

7 REFERENCES

Abebaw, A. (2017). Monitoring the urban growth of Debre Markos Town (1984-2012), Ethiopia: Using satellite images and GPS. Journal of Geography and Regional Planning, 10(4), 69–76. https://doi.org/10.5897/jgrp2016.0533

Agegnehu, S. K., Fuchs, H., Navratil, G., Vuolo, F., & Mansberger, R. (2015). Spatial Urban Expansion and Land Tenure Security in Ethiopia : Case Studies from Bahir Dar and Debre Markos Peri-Urban Areas Spatial Urban Expansion and Land Tenure Security in Ethiopia : Case Studies from Bahir Dar and Debre. USNR, 29(3), 311–328. https://doi.org/10.1080/08941920.2015.1062947

Angel, S. (2023). Urban expansion: theory, evidence and practice. In Buildings and Cities (Vol. 4, Issue 1, pp. 124–138). Web Portal Ubiquity Press. https://doi.org/10.5334/bc.348

Angel, S., Lamson-Hall, P., Blei, A., Shingade, S., & Kumar, S. (2021). Densify and expand: A global analysis of recent urban growth. Sustainability, 13(7), 3835.

Arif, M., & Gill, A. R. (2023). Socio-Economic Determinants of Urbanization in the Perspective of Pakistan. Pakistan Journal of Humanities and Social Sciences, 11(1). https://doi.org/10.52131/pjhss.2023.1101.0387

Ayele, A., & Tarekegn, K. (2020). The impact of urbanization expansion on agricultural land in Ethiopia: A review. Environmental and Socio-Economic Studies, 8(4), 73–80. https://doi.org/10.2478/environ-2020-0024

Braimoh, A. K., & Onishi, T. (2007). Spatial determinants of urban land use change in Lagos, Nigeria. Land Use Policy, 24(2), 502–515. https://doi.org/10.1016/j.landusepol.2006.09.001

Cheng, J., Sun, J., Yao, K., Xu, M., and& Cao, Y. (2022). A variable selection method based on mutual information and variance inflation factor. Spectrochimica Acta -Part A: Molecular and Biomolecular Spectroscopy, 268. https://doi.org/10.1016/j.saa.2021.120652

Chen, Y., Chang, K. T., Han, F., Karacsonyi, D., & Qian, Q. (2016). Investigating urbanization and its spatial determinants in the central districts of Guangzhou, China. Habitat International, 51, 59–69. https://doi.org/10.1016/j.habitatint.2015.10.013

Christensen, P., & McCord, G. C. (2016). Geographic determinants of China's urbanization. Regional Science and Urban Economics, 59, 90–102. https://doi.org/10.1016/j.regsciurbeco.2016.05.001

Dutta, D., Rahman, A., Paul, S. K., & Kundu, A. (2020). Estimating urban growth in peri-urban areas and its interrelationships with built-up density using earth observation datasets. Annals of Regional Science, 65(1), 67–82. https://doi.org/10.1007/s00168-020-00974-8

Fraser, T., Feeley, O., Ridge, A., Cervini, A., Rago, V., Gilmore, K., Worthington, G., & Berliavsky, I. (2024). How far I'll go: Social infrastructure accessibility and proximity in urban neighborhoods. Landscape and Urban Planning, 241. https://doi.org/10.1016/j.landurbplan.2023.104922

Hofmann, A., & Wan, G. (2013). Determinants of Urbanization (355). https://ssrn.com/abstract=2295736

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). Springer Texts in Statistics An Introduction to Statistical Learning with Applications in Python. Springer Nature Switzerland AG.

Jing, S., Yan, Y., Niu, F., & Song, W. (2022). Urban Expansion in China: Spatiotemporal Dynamics and Determinants. Land, 11(3). https://doi.org/10.3390/land11030356

Kmen, C., Navratil, G., & Giannopoulos, I. (2024). Location, Location, Location: The Power of Neighborhoods for Apartment Price Predictions Based on Transaction Data. ISPRS International Journal of Geo-Information, 13(12), 425.

Karimi, F., Sultana, S., Babakan, A. S., & Suthaharan, S. (2021). Urban expansion modeling using an enhanced decision tree algorithm. GeoInformatica, 25(4), 715–731. https://doi.org/10.1007/s10707-019-00377-8Movsessian, A., Cava, D. G., & Tcherniak, D. (2022). Interpretable machine learning in damage detection using shapley additive explanations. ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part B: Mechanical Engineering, 8(2), 021101

Mustafa, A., Heppenstall, A., Omrani, H., Saadi, I., Cools, M., & Teller, J. (2018). Modelling built-up expansion and densification with multinomial logistic regression, cellular automata and genetic algorithm. Computers, Environment and Urban Systems, 67, 147–156. https://doi.org/10.1016/j.compenvurbsys.2017.09.009

Mustafa, A., Van Rompaey, A., Cools, M., Saadi, I., & Teller, J. (2018). Addressing the determinants of built-up expansion and densification processes at the regional scale. Urban Studies, 55(15), 3279–3298. https://doi.org/10.1177/0042098017749176

Nakas, C. T., Bantis, L. E., & Gatsonis, C. A. (2023). ROC Analysis for Classification and Prediction in Practice (1st ed.). CRC Press, Taylor & Francis Group, LLC. https://doi.org/https://doi.org/10.1201/9780429170140

Japkowicz, N. & Shah, M.. (2011). Evaluating Learning Algorithms A Classification Perspective. cambridge university press.

Pesantez-Narvaez, J., Guillen, M., & Alcañiz, M. (2019). Predicting motor insurance claims using telematics data – XGboost versus logistic regression. Risks, 7(2). https://doi.org/10.3390/risks7020070

Sarkar, A., & Chouhan, P. (2020). Modeling spatial determinants of urban expansion of Siliguri a metropolitan city of India using logistic regression. Modeling Earth Systems and Environment, 6(4), 2317–2331. https://doi.org/10.1007/s40808-020-00815-9

Shahri, N. H. N. B. M., Lai, S. B. S., Mohamad, M. B., Rahman, H. A. B. A., & Rambli, A. Bin. (2021). Comparing the performance of adaboost, xgboost, and logistic regression for imbalanced data. Mathematics and Statistics, 9(3), 379–385. https://doi.org/10.13189/ms.2021.090320

Sirko, W., Kashubin, S., Ritter, M., Annkah, A., Salah, Y., Bouchareb, E., Dauphin, Y., Keysers, D., Neumann, M., Cisse, M., & Quinn, J. (2021). Continental-Scale Building Detection from High Resolution Satellite Imagery. In Google Research. https://doi.org/https://doi.org/10.48550/arXiv.2107.12283

Starbuck, C. (2023). The Fundamentals of People Analytics With Applications in R. In The Fundamentals of People Analytics with Applications in R. Springer International Publishing. https://doi.org/10.1007/978-3-031-28674-2

Temesgen, Mekuriaw; & Huseyin, G. (2019). The Impact of Urban Expansion on Physical Environment in Debre Markos Town, Ethiopia. Civil and Environmental Research, August, 16–26. https://doi.org/10.7176/cer/11-5-03

Wu, W., Zhao, S., Zhu, C., & Jiang, J. (2015). A comparative study of urban expansion in Beijing, Tianjin and Shijiazhuang over the past three decades. Landscape and urban planning, 134, 93-106.

Zou, K. H., Liu, A., Bandos, A. I., Ohno-Machado, L., & Rockette, H. E. (2012). Statistical Evaluation of Diagnostic Performance Topics in ROC Analysis (Shein-Chung Chow, Byron Jones, Jen-pei Liu, Karl E. Peace, & Bruce W. Turnbull, Eds.). CRC Press, Taylor & Francis Group.