VOL. 05 NO. 4,
AUGUST, 2024
E-ISSN 3026-9210
P-ISSN 3026-815X

African Scholar Publications & Research International
www.africanscholarpub.com

Journal of Built Env. & Geological Research (JBEGR)

Influence of Geospatial Features on Performance of Machine Learning Models for Predicting Water Consumption by Poor Urban Households

Taiwo, Tolu A.; Olusina J. O.; Hamid-Mosaku A. I.; & Abiodun O. E.

Department of Surveying and Geoinformatics, Faculty of Engineering, University of Lagos, Nigeria.
Corresponding Author: *tolutaiwo75@gmail.com*

Abstract

Several studies have applied various techniques to model and predict water consumption in urban areas, but none has been found that integrates geospatial technology and machine learning techniques to predict water consumption by poor urban households. Using an integrated tool of geospatial technology and machine learning techniques, this study examines the influence of geospatial features on performance of machine learning models for predicting volume of water consumed by poor urban households. Historical data of daily volume of water consumed was gathered through questionnaires, and integrated with socioeconomic data, weather data, property data and geospatial data using geospatial technologies. The datasets were passed through Pearson correlation algorithm to select few features that correlate with the target variable. The selected features were inputted into four predictive models built with four machine learning techniques – Multilinear Regression (MLR), Random Forest (RF), Support Vector Regression (SVR), and Artificial Neural Networks (ANN). Three error metrics, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and R squared (R^2) score were used to measure the model performances. The results show that including GSF considerably improved MLR, RF, SVR and ANN model performances as their RMSE decreased from 139 litres to 116 litres, 77 litres to 57 litres, 130 litres to 110 litres, and 110 litres to 53 litres respectively during training. However, significance test at 95% confidence level carried out on

the results shows that when GSF is included as input into each of the models, improvement in model performance is not significant for MLR, RF and SVR.

Keywords: Water Consumption, Urban Poor, Geospatial Features, Geospatial Technologies, Machine Learning Techniques

Introduction

It is important to model volume of water consumed in a poor urban area before a decision to extend water distribution network (WDN) to the area. Researchers have applied various techniques to model and predict water consumption in urban areas. These techniques include autoregressive integrated moving average (ARIMA) and multilinear regression (MLR) (Adamowski, 2008), artificial neural networks (ANN) (Adamowski, 2012; Ghiassi et al., 2008), random forests and support vector regression (Herrera et al., 2010), recurrent neural network and convolutional neural network (Medina, 2018), and differential-inspired artificial neural network (Oyebode, 2020). Though many studies have been carried out on household water demand in high-income countries, the corresponding body of work for low-income countries is significantly few. Searches done on academic research databases showed a few studies imputing geospatial variables into machine learning models to predict volume of water consumed in high-income countries, but none was found which uses integrated tool of geospatial technology (GST) and machine learning techniques (MLT) to study the influence of GSF on performance of machine learning models. (WSUP, 2016; Casali, et al., 2022; Kopczewska, 2022; Nikparvar & Thill, 2021).

GST provides an environment for processing, analyzing and visualizing geospatial data and how they impact phenomena on, near or beneath the earth surface. Geospatial data includes information about the location, shape, and size of objects, as well as their relationships with other objects in space. Geospatial data or geospatial features (GSF) in machine learning models are variables or geographical features that represent real-world entities that are being modelled. Geospatial features (GSF) are important in machine learning modeling because they provide additional context and information that can improve the accuracy and effectiveness of models. Adding GSF to datasets fed into machine learning models could give insights into the influence of geospatial components on phenomena (Patel, 2021). For example, data about water consumption by the urban poor would usually include socioeconomic, demographic, weather and historical components. Since the

people and their water sources are at definite locations on the earth surface, it means that their activities (water consumption in this case), and data that describe those activities, have geospatial components. The questions: Where? How far apart? And how high? can only be adequately answered if the location is spatially referenced; that is, the location of the phenomena (the people and their water sources) have values from which their distance apart and height difference can be deduced. GSF provides these values. Geospatial components could add to knowledge if explored. For instance, how does distance from the nearest water source influence the per capita water consumption in a household? What is the impact of height difference between the household location and the nearest water source on water consumption? The above example shows the importance of GSF in determining phenomena, such as volume of water consumed in a household.

GSF can provide valuable information for machine learning models (Ioannis, 2019). In predicting water consumption, geospatial data such as land use, distance from water source, and elevation can help identify areas where water is used more or less. Incorporating such information in models can lead to better predictions and more effective management strategies. Furthermore, GSF can help identify spatial dependencies and relationships that may not be apparent from other types of data. For example, water consumption in one location may be affected by activities or conditions in nearby areas. By incorporating GSF, machine learning models can capture these spatial dependencies and improve predictions. GSF can help modelers understand the spatial relationships between different variables. For example, it can be used to identify patterns and trends that occur within specific geographic regions. Geospatial data can be used to create new features that can be used in machine learning models. A model may use the distance and height difference between two locations as features to predict a certain outcome. This information can be used to develop more effective management strategies for water distribution networks.

Machine learning techniques have been applied in the analysis of various geospatial data in many areas of geomatics and the environment such as spatial predictions, classification and mapping (Kang et al., 2021); natural hazards and environmental risk assessments (Mousavi et al., 2011); geoid modelling (Dawod & Abdel-Aziz, 2020); treatment and classification of remote sensing images, assimilation of data and science based models (Kanevski et al., 2008). In various other studies, random forest, K-nearest neighbour and support vector machine were used for land cover classification (Noi & Kappa, 2017; Pal, 2006; Shi & Yang, 2015), spatial data streams (Maleq, 2017), location searching (Gandhimathi et al., 2017), and urban development (Mirbagheri & Alimohammadi, 2018). Additionally, researchers have proposed a combination of neural networks with remote sensing image data for mapping the urbanization dynamics (Li et al., 2019; Liu, 2019; Tayyebi et al., 2011). Multilinear

perceptron has been applied to the classification of remotely sensed data to distinguish land cover types (Benediktsson, 1990; Civco, 1993; Dreyer, 1993; Hepner, 1990; Kanellopoulos et al., 1992; Key et al., 1989; Paola, 1995). Other studies use ANN for classification of land cover and landscape changes on various remote sensing images ((Dreyer, 1993; Foody et al., 1995; Pechanec et al., 2014), for mineral deposit mapping (Rigol-Sanchez et al., 2003), for studying geoid height, coordinate and height transformation (Yilmaz et al., 2017; Ziggah et al., 2019), to predict bus travel times (Arhin, 2021), and to study spatial interpolation (Nevtipilova et al., 2014).

The aim of this study is to examine the influence of geospatial features (GSF) on performance of machine learning models for predicting volume of water consumed by poor urban households. This study is necessary to provide solution to the problem of access to safe water for poor people living within urban areas in low-income countries. Accurately predicting volume of water consumed is the first step towards extending water distribution network to poor people, and solving the problem of lack of access to safe water. Thus, it is necessary to examine how GSF influence model performance.

Methodology

Study Area

Nyanya-Mararaba is a border town between Federal Capital Territory (FCT) and Nasarawa State in Nigeria. It lies between Latitudes 8º24'54''N and 9º18'48''N, and Longitude 6º44'25''E and 7º35'15''E. The town's climate is largely sunny with diurnal temperatures ranging from 18.45 to 36.05 degrees Celsius. Rainfall varies from 0.0 to 400.0 mm/month. The climate is due to the town's location at the transition zone between the 'humid' south and the 'sub-humid' north.

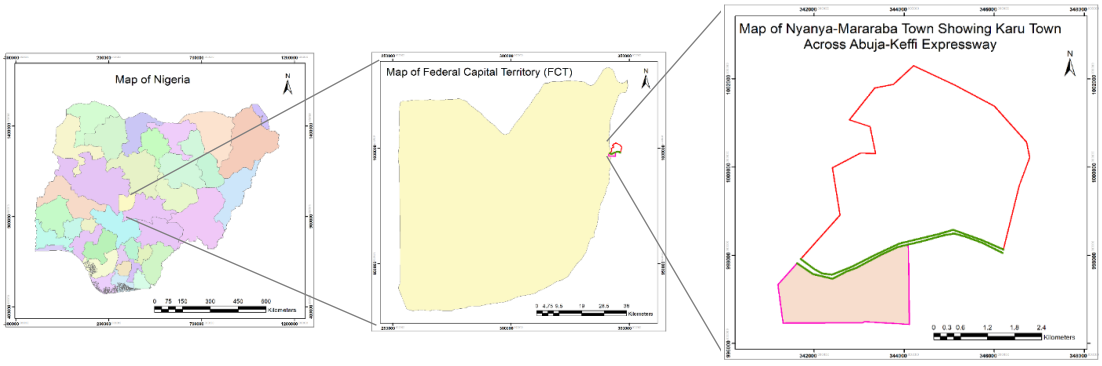


Figure 1: Map of Nigeria and Abuja, showing Nyanya-Mararaba Town

The growing population of the Federal Capital City is concentrated at Nyanya-Mararaba town where an estimated 70% of the people working in the city live. Covering an area of 14 square kilometers, Nyanya-Mararaba is a peri-urban town consisting of informal dwellings where many poor people find shelter.

Methods

Datasets

The datasets and data pre-processing used in this study are described in Taiwo et. al (2023).

Five geospatial data were used in this study: location of households, location of water points, return-trip travel time, elevation and land use land cover type. Geospatial data such as location of households and location of water points were collected with handheld GPS while DEM and LULC data were downloaded from appropriate websites. Location of household and location of water points are represented by the coordinates (shown in Tables 3.2) captured with handheld GPS. Return trip travel time is the time it takes to travel from a household to the nearest water point and back, including queueing time and time taken by detours. It was retrieved from questionnaire responses.

Four GSFs were derived from the geospatial data. They are Travel time, Shortest distance, Height difference, and Land use land cover (LULC) type.

All Features

After feature creation and transformation, there were 30 features all together as shown in Table 1. Volume in lpcd is the dependent variable, the feature to be predicted. Thus, there are 29 independent variables, that is, predictors or explanatory features.

Table 1: All features used in the study

Feature	Type	Description
Volume	Historic	Volume of water consumed in liter per day
ID	Socioeconomic	Identification number of each household
Household income	Socioeconomic	Total income of each household in a month
Education	Socioeconomic	Highest level of education in the household
Household size	Socioeconomic	Number of persons in the household
Rainfall	Weather	Amount of rainfall per month
Ave temp	Weather	Average temperature per month
Travel time	Geospatial	Time it takes to get water and return
Amount spent	Socioeconomic	Amount spent on water per day
Willingness to pay	Socioeconomic	The amount of money a household is willing to pay if it is connected to piped water network
Kitchen Sink	Property	Presence of water sink in the kitchen
ToiletWC	Property	Presence of WC system in the toilet
Garden	Property	Presence of garden in the yard

Car	Property	Ownership of motor vehicle
Shortest distance	Geospatial	Shortest distance to the nearest water point from a household
Height diff	Geospatial	Difference between the household elevation and the nearest water point
Gender_male	Socioeconomic	Gender of respondent – male
Gender_female	Socioeconomic	Gender of respondent – female
Method_carried	Socioeconomic	Method of accessing water – carried from water point
Method_delivered	Socioeconomic	Method of accessing water – delivered to household
Method_inyard	Socioeconomic	Method of accessing water – water point in yard
Method_waterboard	Socioeconomic	Method of accessing water – piped water to the household
Method_well	Socioeconomic	Method of accessing water – well
Availability_not_often	Socioeconomic	Availability of water – not often
Availability_often	Socioeconomic	Availability of water – often
Quality_poor	Socioeconomic	Quality of water – poor
Quality_fair	Socioeconomic	Quality of water – fair
Quality_good	Socioeconomic	Quality of water – good
Quality_very good	Socioeconomic	Quality of water – very good
LULC	Geospatial	Land use land cover type

Feature Selection

Nine features were selected from the experiment carried out by Taiwo et al. (2023). The nine features are household income, household size, rainfall, average temperature, travel time, amount spent, willingness to pay, shortest distance and height difference. These features, listed in Table 2, formed the data input in the machine learning models.

Selected Geospatial Features

Three geospatial features were among the nine features selected. They are travel time, shortest distance and height difference. These make the difference between complete dataset and control dataset used in the experiment.

Table 2: Selected explanatory features

Feature	Symbol
Household income	*I*
Household size	*S*
Rainfall	*R*
Ave Temp	T
Travel Time	*t*

Amount spent	*A*
Willingness to pay	*W*
Shortest Distance	*d*
Height difference	*h*

Complete Dataset

After feature creation and transformation, there were 30 features all together as shown in Table 1. Volume in liter per day is the dependent variable, the feature to be predicted. Thus, there are 29 predictor variables (listed in Table 1), together with 1200 observations, they form the complete dataset.

Control Dataset

The complete dataset without geospatial features acts as the control dataset. The aim is to use the control dataset to investigate the importance of geospatial features in the performance of machine learning models for predicting water consumption by poor urban households.

Coding the Models

Multilinear Regression (MLR), Random Forest (RF), Support Vector Regression (SVR), and Artificial Neural Network (ANN) were chosen for this work because they represent machine learning techniques in the supervised learning category for regression analysis (Kiwelekar et al., 2020).

The complete dataset was first one-hot encoded before it was split into training, validation and test datasets. The MLR, RF, SVR and ANN models, which were named *model* in the codes, were each created as an instance of their respective classes in Scikit Learn® library: *SGDRegressor, RandomForestRegressor, LinearSVR and MLPRegressor*, which were each infused with a scaler to scale the dataset (Pedregosa et al., 2012). *model* was fitted to the training dataset in order to learn the data. Three functions named *training, validation* and *testing* were coded to carry out prediction with training dataset, validation dataset and test dataset respectively. Each of them calls another function named *evaluation*, which evaluates the prediction with the three evaluation metrics and prints the results. The predicted volume of water consumed was printed into a csv file.

Multilinear Regression

MLR models actual and predicted volume of water consumption using equations (1) and (2).

$$Y^{o(i)} = \theta_1 I^{(i)} + \theta_2 S^{(i)} + \theta_3 R^{(i)} + \theta_4 T^{(i)} + \theta_5 t^{(i)} + \theta_6 A^{(i)} + \theta_7 W^{(i)} + \theta_8 d^{(i)} + \theta_9 h^{(i)} + \boldsymbol{\theta_o} \quad (1)$$

$$Y^{p(i)} = \theta_1 I^{(i)} + \theta_2 S^{(i)} + \theta_3 R^{(i)} + \theta_4 T^{(i)} + \theta_5 t^{(i)} + \theta_6 A^{(i)} + \theta_7 W^{(i)} + \theta_8 d^{(i)} + \theta_9 h^{(i)} + \boldsymbol{\theta_o} \quad (2)$$

where i = (1, …, 1200) for 1200 examples or households in this case,
Y^o = Actual volume of water consumed by a household,
Y^p = Predicted volume of water consumed by a household,
θ = parameters, θ_o = bias or y-intercept, and
I, S, R, T, t, A, W, d, h = selected explanatory variables defined in Table 1 and Table 2.

Equation (3) is the system of equations of the actual volume for 1200 examples or households:

$$Y^{o(1)} = \theta_1 I^{(1)} + \theta_2 S^{(1)} + \theta_3 R^{(1)} + \theta_4 T^{(1)} + \theta_5 t^{(1)} + \theta_6 A^{(1)} + \theta_7 W^{(1)} + \theta_8 d^{(1)} + \theta_9 h^{(1)} + \boldsymbol{\theta_o}$$
$$Y^{o(2)} = \theta_1 I^{(2)} + \theta_2 S^{(2)} + \theta_3 R^{(2)} + \theta_4 T^{(2)} + \theta_5 t^{(2)} + \theta_6 A^{(2)} + \theta_7 W^{(2)} + \theta_8 d^{(2)} + \theta_9 h^{(2)} + \boldsymbol{\theta_o}$$
$$\vdots$$
$$Y^{o(1200)} = \theta_1 I^{(1200)} + \theta_2 S^{(1200)} + \theta_3 R^{(1200)} + \theta_4 T^{(1200)} + \theta_5 t^{(1200)} + \theta_6 A^{(1200)} + \theta_7 W^{(1200)} + \theta_8 d^{(1200)} + \theta_9 h^{(1200)} + \boldsymbol{\theta_o} \quad (3)$$

In matrix form, $Y = A\theta + \theta_o$ (4)

We define a cost function (the objective function; also called squared error function), which is the sum of squares of the residuals, expressed in equation (5):

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}(Y^{p(i)} - Y^{o(i)})^2 \quad (5)$$

Given a training examples $(x_1, y_1), \ldots, (x_m, y_m)$, we want to learn a linear function $A\theta + \theta_o$ with model parameters θ and intercept θ_o. To find the model parameters, we minimize the cost function given by equation (6).

$$Min_{\theta_0,\ldots\theta_9} \frac{1}{m}\sum_{i=1}^{m}(Y^{p(i)} - Y^{o(i)})^2 \quad (6)$$

MLR is similar to ordinary least squares method, which can also be used to estimate the parameters. The ordinary least squares method seeks to minimize the sum of the

squared residuals. As in most linear regression models, the objective is to minimize the objective function, as expressed by equation (6). MLR, on the other hand, trains learning algorithms using labelled training data to understand the relationship between features and an outcome. In MLR, equation (6) is optimized by Stochastic Gradient Descent (SGD), an optimization algorithm for unconstrained optimization problems. SGD approximates the true gradient of the objective function by considering a single training example at a time. It minimizes the cost of predicting outcomes after fitting the model and gives the values of the parameters w that produce the minimum cost.
For each value of parameter θ, we derive a value for $J(\theta)$ from equation 3.8. Plotting different values of $J(\theta)$ against values of parameter w gives a curve, whose global minimum corresponds to the value of parameter θ that minimizes the cost function $J(\theta)$. The line that fits the data well is the line of prediction given by the value of parameter w that minimizes $J(\theta)$.
Method of normal equation could be used to do similar computation (Ayeni, 1983; Ghilani, 2018), but SGD algorithm scales better for large datasets (Ng, 2013). The normal equation is a closed-form solution of computing the value of parameter $\boldsymbol{\theta}$ that minimizes the cost function. Ghilani (2018) gives the normal equation as equation (7),

$$A^T A\theta = \theta^T Y \tag{7}$$

where A is the matrix of coefficients or design matrix, and the solution as equation (8).

$$\theta = (A^T A)^{-1} A^T Y \tag{8}$$

Equation (8) gives the best parameters θ that minimizes the cost function. To make prediction with the parameters θ, we use equation (9).

$$Y^p = \theta^T A \tag{9}$$

The normal equation was hard-coded as a function named *normal_equation*. The function calls the *evaluation*, function, which outputs the evaluation metrics MAE, RMSE and R^2 each time *normal_equation* is called. The results from the machine learning models and normal equation model were compared in this study.

Random Forest

The same coding method used for MLR model was employed for RF, except that no optimizer was used on RF. Optimizers are not directly used in RF regression as RF does not involve an explicit optimization process to find the best hyper parameters or

tune the model. To find the optimal values of its hyper parameters, grid search technique was used. This method systematically explores different combinations of hyper parameter values and evaluate the model's performance on a validation set. Three hyperparameters are explored and their values are stated in Table 3.

Table 3: RF hyper parameters, their descriptions and values assigned to them

Parameters	Description	Values assigned
random_state	Controls randomness of bootstrapping and sampling of features	42
n_estimators	The number of trees in the forest	100
max_depth	The maximum depth of the tree	5

Support Vector Regression

Same coding method was also employed to develop SVR model without an optimizer. The primary focus SVR is finding the best hyperplane that fits the training data within a specified margin. The two hyper parameters and their values that describe the best hyperplane are C and epsilon given in Table 4.

Table 4: SVR hyper parameters, their descriptions and values assigned to them

Parameters	Description	Values assigned
C	error penalty	1.0
epsilon	slack variable	0.2

Artificial Neural Network

Coding ANN is almost like coding MLR except that ANN carries along a number of parameters as shown in Table 5. The table lists the hyper parameters and the values assigned after tuning the parameters. ANN is optimized by Adaptive Momentum (Adam) a further extension of stochastic gradient descent that updates network weights during training.

Table 5: ANN hyper parameters, their descriptions and values assigned to them

Parameters	Description	Values assigned
hidden_layer_sizes	Number of hidden layers	32,
activation	Activation function	'relu'
solver	Optimizer	'adam'
learning_rate_init	Learning rate	0.01
max_iter	Maximum number of iterations	2000

Coding Model Evaluation

Three error metrics were used to measure the performance of the model: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and R^2 score. MAE is used

to expressed in equation (10), to measure the absolute difference between observed and predicted values of the target variable.

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|Y_i^o - Y_i^p| \quad (10)$$

Mean Square Error (MSE) more greatly penalizes larger errors and is sensitive to outliers due to taking the squared difference between variables (Mueller, 2021). Also, MSE is good for values close to zero but not for large values. We want both large and small errors to be penalized equally so that we could obtain a robust model that would predict well. Therefore, we used root mean squared error (RMSE), which is the square root of the value obtained from MSE, expressed in equation (11). It shows how far predictions fall from actual values. The smaller the RMSE the closer the fit to the data.

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(Y_i^o - Y_i^p)^2} \quad (11)$$

R^2 score is the amount of the variation in the target variable which is predictable from the explanatory variables as mathematically shown in equation (12). It is used to check how well observed results are reproduced by the model. Best possible score is 1.0, but the scores are stated as percentages in this study.

$$R^2 = \left[\frac{\sum_{i=1}^{m}(Y_i^o - \bar{Y}_i^0)\left(Y_i^p - \bar{Y}_i^p\right)}{\sqrt{\sum_{i=1}^{m}\left(Y_i^o - \bar{Y}_i^0\right)^2 . \sum_{i=1}^{m}\left(Y_i^p - \bar{Y}_i^p\right)^2}}\right]^2 \quad (12)$$

In the three expressions above, Y_i^o, Y_i^p, $\bar{Y}_i^o$, and $\bar{Y}_i^p$ are observed and predicted values and their respective average values; m is the number of observations or examples. The three error metrics were hard-coded with Python codes in Jupyter Notebook®.

Validation

The models were validated in two ways: firstly, they were validated with the validation dataset during modelling. Secondly, the models were validated by comparing their performances with the performances of normal equation.

Results and Discussion

Examining Influence of Geospatial Features (GSF) on Model Performance

Using the complete dataset and control dataset described above, experiments were carried out to investigate the influence of GSF on model performances. Table 6, Table

7 and Table 8 show model performance with GSF (using the complete dataset) and without GSF (using the control dataset) included in the explanatory variables. The results show that including GSF considerably improved Multilinear Regression model performance as RMSE value decreased from 139 litres to 116 litres and R^2 score rose from 50% to 65% during training. Also, there is improvement in Random Forest model performance as RMSE value decreased from 77 litres to 57 litres and R^2 score increases from 85% to 92% during model training. Similarly, when GSF was added, Support Vector Regression model performance decreased from 130 litres to 110 litres and R^2 score rose from 58% to 70% during training. Artificial Neural Networks model also shows performance improvement with the addition of GSF as RMSE value decreased from 110 lires to 53 litres and R^2 score increases from 68% to 93% during training. Figure 2 shows the plots of predicted volume against actual volume of water consumed when GSF were added to the variables and when they were not. All four models performed better when GSF added into them.

Table 6: Model performances with GSF and without GSF during TRAINING

Model	With GSF			Without GSF		
	MAE (litre)	**RMSE (litre)**	**R^2 (%)**	**MAE (litre)**	**RMSE (litre)**	**R^2 (%)**
Multilinear Regression	104	116	65	116	139	50
Random Forest	41	57	92	55	77	85
Support Vector Regression	90	110	70	103	130	58
Artificial Neural Network	40	53	93	83	110	68
Normal Equation	86	105	71	98	123	59

Table 7: Model performances with GSF and without GSF during VALIDATION

Model	With GSF			Without GSF		
	MAE (litre)	**RMSE (litre)**	**R^2 (%)**	**MAE (litre)**	**RMSE (litre)**	**R^2 (%)**
Multilinear Regression	113	124	54	102	116	57
Random Forest	46	62	91	59	86	80
Support Vector Regression	91	116	72	100	127	63
Artificial Neural Network	45	61	90	89	112	70
Normal Equation	78	97	76	104	133	54

Table 8: Model performances with GSF and without GSF during TESTING

Model	With GSF			Without GSF		
	MAE (litre)	**RMSE (litre)**	**R^2 (%)**	**MAE (litre)**	**RMSE (litre)**	**R^2 (%)**
Multilinear Regression	97	109	69	109	144	42
Random Forest	45	60	89	61	87	79
Support Vector Regression	67	107	73	99	132	58
Artificial Neural Network	43	57	92	84	112	61
Normal Equation	84	103	72	95	120	63

Significance Test on Difference in Model Performances with GSF and without GSF
The null hypothesis states that the difference between model performances when GSF was included and when GSF was not included is not significant. The p value of each model is stated in Table 9. Since the p value is less than 0.05 for MLR, RF and ANN, except SVR, the null hypothesis is accepted; that is, there is no significant difference in model performance when GSF was included and when GSF was not included in the modelling.

Null hypotheses: $H_0 : \mu = \mu_0$, that is, there is no significant difference in model performance with GSF and without GSF

Alternative hypothesis: $H_a : \mu \neq \mu_0$, that is, there is significant difference in model performance with GSF and without GSF.

Table 9: Two-tails t-test at 95% confidence level

Model	p **value**	**Degrees of Freedom**	**Test Result**
Multilinear Regression	0.5756	8	$H_0 : \mu = \mu_0$
Random Forest	0.2093	8	$H_0 : \mu = \mu_0$
Support Vector Regression	0.0007	8	$H_a : \mu \neq \mu_0$
Artificial Neural Network	0.3397	8	$H_0 : \mu = \mu_0$

Model Validation with Normal Equation
The machine learning models were validated with the performance of normal equation. The performance of normal equation is also shown in Table 6, Table 7 and Table 8. Though Random Forest and Artificial Neural Network outperformed Normal

Equation, it also performed considerably well with RMSE value of 58 litres and R^2 score of 88% during training.

Significance Test on Difference in Model Performance Between Normal Equation and Machine Learning Models

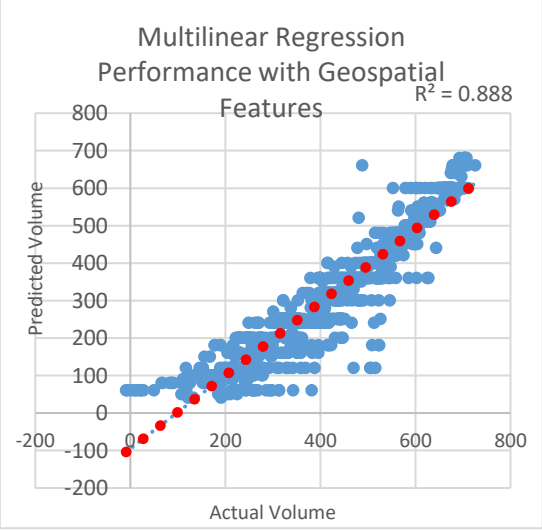
To validate the performance of the machine learning models, significance test was carried out with t-test at 95% confidence level on the results produced by normal equation and machine learning models. The null hypothesis states that there is no significant difference between normal equation model performance and machine learning model performance. The p value of each model is stated in Table 10. Since the p value is more than 0.05 for MLR, RF and SVR except ANN, there is a reason to accept null hypothesis; that is, there is no significant difference between normal equation model performance and machine learning model performance.

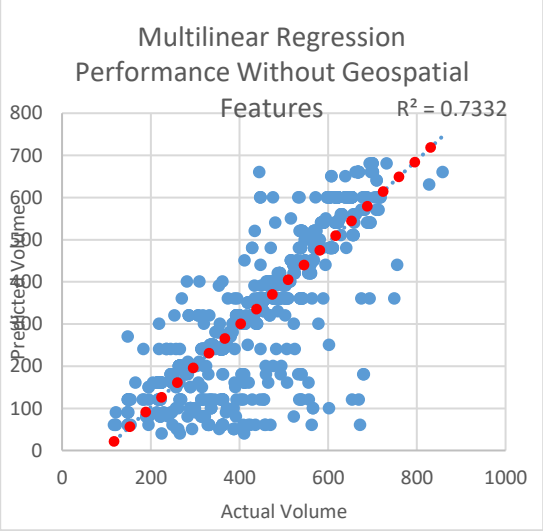
Null hypothesis: $H_0 : \mu = \mu_0$, that is, there is no significant difference between normal equation model performance and machine learning model performance.

Alternative hypothesis: $H_a : \mu \neq \mu_0$, that is, there is significant difference between normal equation model performance and machine learning model performance.

Table 10: Two-tailed t-test at 95% confidence level

Model	*p* value	Degrees of freedom	Test Result
Multilinear Regression	0.5049	8	$H_0 : \mu = \mu_0$
Random Forest	0.0147	8	$H_0 : \mu = \mu_0$
Support Vector Regression	0.1469	8	$H_0 : \mu = \mu_0$
Artificial Neural Network	0.0026	8	$H_a : \mu \neq \mu_0$





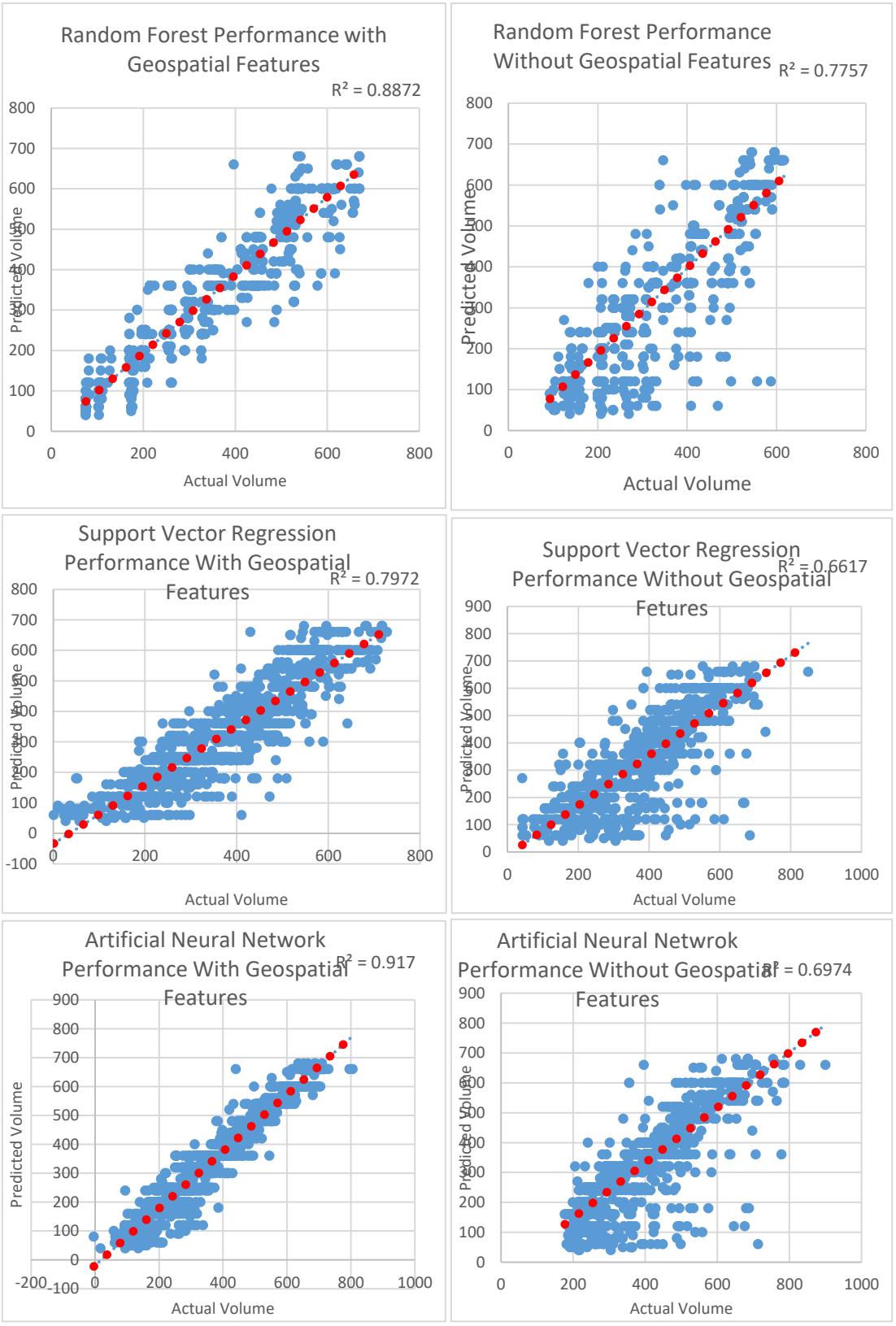


Figure 2: Actual and predicted volume of water consumed with GSF and without GSF

Conclusion

This study examined the influence of geospatial features (GSF) on performance of machine learning models for predicting volume of water consumed by poor urban households. Three GSF – travel time, shortest distance and height difference, which correlated with the target variable – were selected from 29 features. The influence of the GSF on Multilinear Regression, Random Forest, Support Vector Regression and Artificial Neural Networks models were examined. The results show that including GSF considerably improved Multilinear Regression model performance as RMSE value decreased from 139 litres to 116 litres and R^2 score rose from 50% to 65% during training. Also, for Random Forest model performance, RMSE value decreased from 77 litres to 57 litres and R^2 score increases from 85% to 92% during model training, while that for Support Vector Regression model performance decreased from 130 litres to 110 litres and R^2 score rose from 58% to 70%. Artificial Neural Networks model also shows performance improvement with the addition of GSF as RMSE value decreased from 110 lires to 53 litres and R^2 score increases from 68% to 93%, during training. Though adding GSF to the models show improvement in model performance but t-test carried out on the results at 95% confidence level shows the significance of the improvement. The test shows that including GSF in Multilinear Regression, Random Forest and Support Vector Regression has no significant difference on model performance. But the improvement is significant for Artificial Neural Network. These results show the importance of including GSF to machine learning models for predicting volume of water consumed by poor urban households.

References

Adamowski Jan, Shiv O. Prasher, Bogdan Ozga-Zielinski, Anna Sliusarieva. 2012. Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in montreal, canada. *Water Resources Research, 48*(1).

Adamowski, J. F. 2008. Peak Daily Water Demand Forecast Modeling Using Artificial Neural Networks. *Journal of Water Resources Planning and Management, 134*(2), 119-128. doi:10.1061/(asce)0733-9496(2008)134:2(119)

Arhin Stephen, B. M., Hamdiat Baba-Adam Adam Gatiba. 2021. Predicting Bus Travel Times in Washington, DC Using Artificial Neural Networks (ANNs). . *Mineta Transportation Institute.* doi:DOI: 10.31979/mti.2021.19437

Benediktsson, J. A., Swain, P.H., Ersoy, O.K. 1990. Neural Network Approaches versus Statistical Methods in Classification of Multisource Remote Sensing Data. *IEEE Transactions on Geoscience and Remote Sensing,, 28*, 540-551.

Casali, Y., Aydin, N. Y., Comes, T. 2022. Machine learning for spatial analyses in urban areas: a scoping review. *Sustainable Cities and Society, 85*, 104050. doi:https://doi.org/10.1016/j.scs.2022.104050

Civco, D. L. 1993. Artificial neural networks for land-cover classification and mapping. *International Journal of Geographical Information Systems, 7*(2), 173-186. doi:10.1080/02693799308901949

Colburn, A. T., Johnson, E. C., Péronnet, F., Jansen, L. T., Capitan-Jimenez, C., Adams, J. D., Kavouras, S. A. 2021. Validity and Reliability of a Water Frequency Questionnaire to Estimate Daily Total Water Intake in Adults. *8*. doi:10.3389/fnut.2021.676697

Dawod, G. M., Abdel-Aziz, T. M. 2020. Utilization of geographically weighted regression for geoid modelling in Egypt. *Journal of Applied Geodesy, 14*(1), 1-12. doi:10.1515/jag-2019-0009

Dreyer, P. 1993. Classification of land cover using optimized neural nets on SPOT data. *Photogrammetric Engineering Remote Sensing, 59*, 617-621.

Fan, L., Wang, F., Liu, G., Yang, X., Qin, W. 2014. Public Perception of Water Consumption and Its Effects on Water Conservation Behavior. *6*(6), 1771-1784.

Foody, G. M., Mcculloch, M. B., Yates, W. B. 1995. Classification of remotely sensed data by an artificial neural network: issues related to training data characteristics. *Photogrammetric Engineering Remote Sensing, 61*, 391-401.

Gandhimathi, D., Kanmani, K., Kirupa, S. 2017. K-NN Process In Applications Of GIS Technologies Using Spatial Temporal Data. *IOSR Journal of Computer Engineering*(Next Generation Computing Technologies.

Ghiassi, M., Zimbra, D., Saidane, H. 2008. Urban Water Demand Forecasting with a Dynamic Artificial Neural Network Model. *Journal of Water Resources Planning Management, 134*, 138-146.
Goodfellow Ian, Y. B., Aaron Courville. 2016) *Deep Learning*.
Hepner GF, L. T., Ritter N, Bryant N. 1990. ANN classification using a minimal training set: comparison to conventional supervised classification. *Photogrammetric Engineering & Remote Sensing, 56*, 469–473.
Herrera, M., Lu´ıs, T., Joaqu´ın, I., Rafael, P. 2010. Predictive models for forecasting hourly urban water demand. Journal of hydrology. *387(1-2):141– 150, 2010., 387(1-2)*, 141– 150.
Kanellopoulos, I., Varfis, A., Wilkinson, G. G., Mégier, J. 1992. Land-cover discrimination in SPOT HRV imagery using an artificial neural network - a 20-class experiment. *13*, 917-924.
Ioannis, D. 2019. Re-Discovering "Big Data" and "Data Science" *in Geodesy and Geomatics. 4th Joint International Symposium on Deformation Monitoring (JISDM), Athens, Greece*.
Kanevski, M., Pozdnoukhov, A., Timonin, V. 2008. Machine learning algorithms for geo spatial data. Applications and software tools. *1*, 320-327.
Kang, Y., Nam, J., Kim, Y., Lee, S., Seong, D., Jang, S., Ryu, C. 2021. Assessment of Regression Models for Predicting Rice Yield and Protein Content Using Unmanned Aerial Vehicle-Based Multispectral Imagery. *13*(8), 1508.
Key, J. R., Maslanik, J., Schweiger, A. J. J. P. E., Sensing, R. 1989. Classification of merged AVHRR and SMMR Arctic data with neural networks. *55*, 1331-1338.
Kiwelekar, A. W., Mahamunkar, G. S., Netak, L. D., Nikam, V. B. 2020. Deep learning techniques for geospatial data analysis. *Machine Learning Paradigms: Advances in Deep Learning-based Technological Applications*, 63-81.
Kopczewska, K. 2022. Spatial machine learning: new opportunities for regional science. *The Annals of Regional Science, 68*(3), 713-755. doi:10.1007/s00168-021-01101-x
Li, J., Huang, X., Gong, J. 2019. Deep neural network for remote-sensing image interpretation: status and perspectives. *National Science Review, 6*, 1082-1086. doi:10.1093/nsr/nwz058
Liu, X. H., F. Ghazali, K.H. Mohamed, I.I. Zhao, Y. 2019. *A review of Convolutional Neural Networks in Remote Sensing Image.* Paper presented at the 8th International Conference on Software and Computer Applications.
Maleq, K. 2017. *K-Nearest Neighbor Classification on Spatial Data Streams Using P-Trees.* Paper presented at the 6th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining. .
Mangal, A., Holm, E.A. 2018. Applied machine learning to predict stress hotspots II: Hexagonal close packed materials. *International Journal of Plasticity.*
Medina, I. 2018. *Predicting Short-Term Water Consumption for Multi-Family Residences.* (Master of Science), University of Ontario, Canada.,
Mirbagheri, B., & Alimohammadi, A. 2018. Integration of Local and Global Support Vector Machines to Improve Urban Growth Modelling. *7*(9), 347.
Mousavi, S. Z., Kavian, A., Soleimani, K., Shirzadi, A. 2011. GIS-based spatial prediction of landslide susceptibility using logistic regression model. *Journal of Geomatics, Natural Hazards and Risk, 2*(1), 33-50. doi:10.1080/19475705.2010.532975
Mueller, J. P., Massaron, L. 2021. *Machine Learning for Dummies*® John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030-5774, Canada.
Nevtipilova, V., Pastwa, J., Boori, M. S., Vozenilek, V. 2014. Testing artificial neural network (ANN) for spatial interpolation. *Journal of Geology Geophysics, 3*(2), 01-09.
Ng, A. 2013. *Lecture Notes. Stanford University*. Retrieved from https://www.datasciencecentral.com/lecture-notes-by-ng-full-set/ accessed on 26th July, 2023.
Nikparvar, B., Thill, J. C. 2021. Machine Learning of Spatial Data. *10*(9), 600.
Noi, T., Kappa, M. 2017. Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery. *18*(1).
Oyebode, O. K. 2020. *Development of A Sustainable Evolutionary Inspired Artificial Intelligent System for Municipal Water Demand Modelling.* (Doctor of Philosophy in Civil Engineering), University of KwaZulu-Natal, South Africa.
Pal, M. 2006. Support vector machine-based feature selection for land cover classification: a case study with DAIS hyperspectral data. *International Journal of Remote Sensing, 27*(14), 2877-2894.
Paola, J. D. a. S., Robert A. 1995. A detailed comparison of backpropagation neural network and maximum-likelihood classifiers for urban land use classification. *IEEE Transactions on Geoscience, 33*(4), 981-996.
Patel, H. 2021. What is Feature Engineering — Importance, Tools and Techniques for Machine Learning. Retrieved from https://towardsdatascience.com/what-is-feature-engineering-importance-tools-and-techniques-for-machine-learning on 26th July, 2023.
Pechanec, V., Burian, J., Dobesova, Z. 2014. Integrating neural networks and GIS in modelling landscape changes. *14th SGEM GeoConference on Informatics, Geoinformatics, 1*, 651-658.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Louppe, G. 2012. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*.
Pragati, B. 2023. Train Test Validation Split: How To and Best Practices. Retrieved from https://www.v7labs.com/blog/train-validation-test-set
Rigol-Sanchez, J., Chica-Olmo, M., Abarca-Hernández, F. 2003. Artificial neural networks as a tool for mineral potential mapping with GIS. *International Journal of Remote Sensing - INT J REMOTE SENS, 24*, 1151-1156. doi:10.1080/0143116021000031791
Säve-Söderbergh, M., Toljander, J., Mattisson, I., Åkesson, A., Simonsson, M. 2017. Drinking water consumption patterns among adults-SMS as a novel tool for collection of repeated self-reported water consumption. *Journal of exposure science & environmental epidemiology, 28*. doi:10.1038/jes.2017.8
Schober P., Boer C. and Schwarte L. A., 2018. Correlation Coefficients: Appropriate Use and
Interpretation. *Anesthesia & Analgesia.* DOI: 10.1213/ANE.0000000000002864

Shi, D., & Yang, X. 2015. Support vector machines for land cover mapping from remote sensor imagery. *Monitoring Modeling of Global Changes: A Geomatics Perspective*, 265-279.

Taiwo, Tolu A. O., Olusina J.O., Hamid-Mosaku A. I., Abiodun O. E. 2023. An

Investigation of the Performance of Geospatial Features in Machine Learning Feature Selection Techniques. *Nigerian Journal of Environmental Sciences and Technology*, 7(2), pp. 277-290. https://doi.org/10.36263/nijest.2023.02.0422 or https://nijest.com/current-issue/

Tayyebi, A., Pijanowski, B. C., Tayyebi, A. H. 2011. An urban growth boundary model using neural networks, GIS and radial parameterization: An application to Tehran, Iran. *Landscape Urban Planning, 100*(1-2), 35-44.

WGF 2012. Human Rights-Based Approaches and Managing Water Resources: Exploring

the potential for enhancing development outcomes. WGF Report No. 1. Stockholm, Stockholm International Water Institute (SIWI).

WSUP 2016. Modelling domestic water consumption in Nairobi and Accra. *Water and Sanitation for Urban Programme.* University of Leeds.

Yee, H. J., Ting, C. Y., Ho, C. C. 2018. Optimal geospatial features for sales analytics. In *AIP Conference Proceedings* (Vol. 2016, No. 1). AIP Publishing.

Yilmaz, M., Turgut, B., Gullu, M., Yilmaz, I. 2017. Application of artificial neural networks to height transformation. *Tehnicki Vjesnik, 24*, 443-448. doi:10.17559/TV-20151116094353

Ziggah, Y. Y., Youjian, H., Tierra, A. R., Laari, P. B. 2019. Coordinate transformation between global and local data based on artificial neural network with k-fold cross-validation in Ghana. *Earth Sciences Research Journal, 23*(1), 67-77.