An Approach Of Urban Expansion Prediction Using Svm And Var

Kazi Abdul Mohite, Dr. Mohammad Rezwanul Huq

Abstract: Chittagong city corporation, a part of the Chittagong district, which is considered as a major business area of Bangladesh, has experienced immense expansion for the proliferation of numerous business activities and unplanned urbanization. The main objective of the paper is to analyze and predict the urban expansion to determine the land requirement for the habitation of people in the future. Nowadays, many machine learning techniques are used for the classification of built-up and non-built up areas. In this paper, various remote sensing data freely available for education and research from 2013 to 2019 are used as data sources. Likewise, the Maximum Likelihood Event (MLE), a supervised classification is used for land cover classification, which is extracted along with the other 9 most influencing factors. Afterward, the Support Vector Machine (SVM) model helps classify the data based on those factors. Besides, Vector Auto-regression (VAR) is also developed as a multivariate forecasting algorithm to forecast the growth of the built-up area in 2021. The data labeled as built-up by the SVM classifier are used to train the VAR model. The result indicates the SVM model is very effective classifying the data when no outliers are present while the VAR model can forecast when a time series data is stationary and correlated with the past time series of other variables. These two models together are highly efficient for explaining urban expansion from the relationship of those factors and their behavior.

Index Terms: Urbanization, Support Vector Machine (SVM), Vector Auto Regression (VAR), Remote sensing, GIS, Machine Learning, Land Cover Change.

1 INTRODUCTION

Urbanization, which is the transformation of other types of land from natural to artificial for the growth of population and economic activities, is a main influencing type of land use and land cover change. It has many detrimental impacts on the climate. Moreover, if rampant urbanization continues, it dwindles arable land and declines food production. Furthermore, massive urbanization could lead to the use of fertile land for industrialization which brings more pollution and proliferates slams around cities. By covering with impervious surfaces, roads, and buildings, urban areas tend to experience higher temperatures compared with the surrounding rural areas. Moreover, urban areas are built mostly on lower elevation and slop since hills and forests are destructed indiscriminately to accommodate a growing population and industries. Consequently, this phenomenon of urbanization could be used for finding out the influencing variables and their future impacts. Land Covers, which are the major sources of information about how land is being used, hold most of the material and energy movement and interaction between the geosphere and biosphere. Hence, land use and land cover change have many environmental effects on a local and regional level and are linked with the global environmental process. Moreover, the change of one element could affect the others for the correlated nature of elements of the environment. Therefore, it is imperative to study the change which provides knowledge about soil quality, runoff, sedimentation rate, biodiversity and the transformation of land-use change for human intervention. The study of the urbanization process in computer science is different from other disciplines since it does not allow physical examination of land.

- *Kazi Abdul Mohite is currently pursuing master's degree program in Computer Science and Engineering in East West University, Bangladesh, PH-8801925200267. E-mail: mohitekazia@gmail.com.*
- *Dr. Mohammad Rezwanul Huq is currently working as Assistant Professor in Computer Science and Engineering in East West University, Bangladesh, PH-8801729648344. E-mail: mrhuq@ewubd.edu*

Thus, artificial simulation and model building methodologies are used for identifying the complexity of land-use change dynamics. Nowadays, the integration of remote sensing and geographic information system have been widely applied and recognized as one of the efficient tools for identifying urban land-use and cover change. Satellite remote sensing collects geographical resources from different elevation and converts them into valuable information for monitoring land-use change and building land cover datasets. The complexity of human activities and the influence of various spatial and temporal dynamics in the urbanization process make the studies of spatial and temporal parameters and their effect inevitable. Many researchers have already used many theoretical and practical modeling techniques to understand correlated variables that drive urbanization. One of the most sophisticated techniques is the Support vector machine (SVM) model, which provides a clear picture of integral factors affecting urbanization. The SVM allows to include many socioeconomic and demographic factors along with minimizing misclassification. Moreover, the use of Vector Auto Regression (VAR) model alongside the SVM model can also enable forecasting the future growth. This literature reviews the SVM and VAR model approach to model land use and cover change and predicts the urban growth.

2 STUDY AREA

This study has been conducted in Chittagong City Corporation. Fig. 1. shows Chittagong city corporation as the study area of the paper. It has an area of 160.99 square km, located on the bank of the Karnaphuli river in between 22°13' and 22°27' north latitudes and in between 91°40' and 91°53' east longitudes. The geography of its area consists of a lot of hills with elevation ranging from 5 to 10 m above mean sea level. The city has a tropical monsoon climate and enjoys a warm and humid climate where temperature varies between 26 °C and 36 °C and average humidity of 78%. Karnaphuli River, which passes through the city, is used for transporting goods for local areas to neighbor cities. In the rainy season, water overflows the river and brings sediment which ensures a huge production of cash crops.

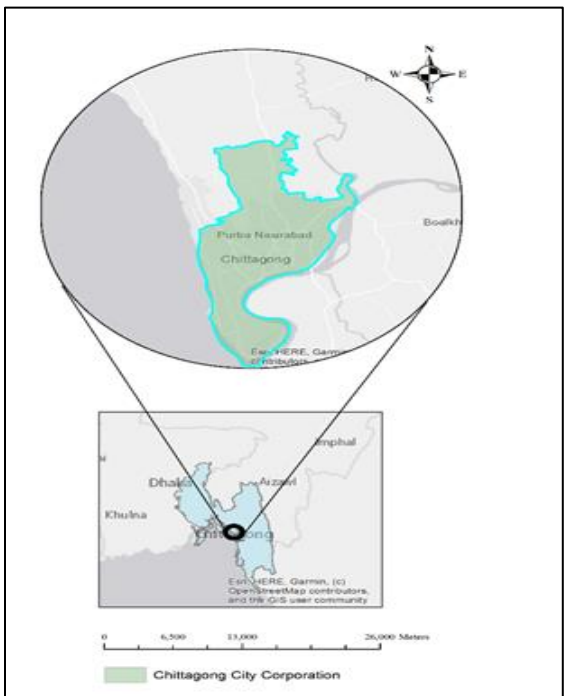

Fig. 1. The study area (Chittagong City Corporation).

3 METHODOLOGY

Urban expansion changes vacant or fertile lands into built-up lands. The expansion depends on various correlated social, economic, and environmental factors. This relationship between these factors and their influence in the change can be described by urban expansion model. Fig. 2. shows an urban expansion prediction model based on the SVM, which classifies the data between built-up and non-built up areas. Afterward, the VAR model uses built-up labeled data to train and forecast for the year 2021. The study considers, a sample data could turn into built-up area in the future if it belongs to that forecasted range and is already classified as non-built up area.

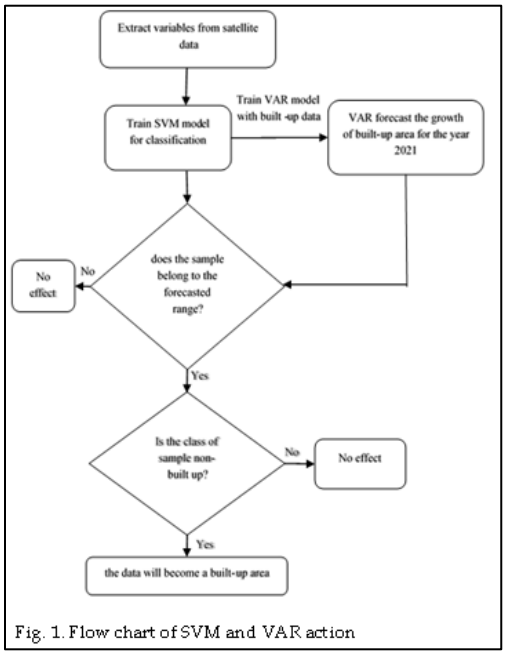

Fig. 1. Flow chart of SVM and VAR action

3.1 A Background of SVM Based Model

SVM is a very strong tool which decides boundary between classes to enable the prediction of labels from one or more feature vectors. The decision boundary known as hyperplane is to build maintaining the best possible distance from the closest data points of each class. Fig. 3. shows an SVM model representation with a hyperplane. The closest points are called Support Vectors. Given a labeled training dataset

$$(x_1, y_1), \ldots, (x_n, y_n), x_i \in R_d \text{ and } y_i \in (-1, +1) \quad (1)$$

In Equation (1) x_1 and y_1 is a feature vector and y_i is the class label (negative or positive) representation of a training dataset i. Then, the optimal hyperplane can be defined as:

$$wx^T + b = 0 \quad (2)$$

In Equation (2) x is the input feature vector, w is the weight vector and b are the bias. The w and b need to satisfy the following inequalities in equation (3) and (4) for all elements of the training dataset.

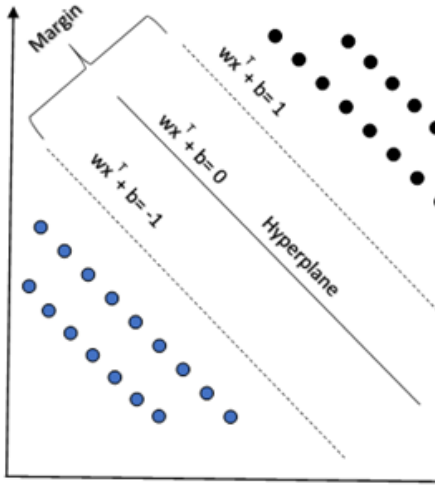

Fig. 1. An SVM model

(3)

$$wx_i^T + b \geq +1 \text{ if } y_i = 1$$

$$wx_i^T + b \leq -1 \text{ if } y_i = —1 \quad (4)$$

The objective of training an SVM model is to find out w and b so that the hyperplane separates the data and maximizes the margin $^{b}/_{||w||}$.

3.2 Introducing the VAR

Vector Auto Regression (VAR) was introduced by Sims (1980). The joint dynamic behavior of a collection of variables without requiring strong restrictions of the kind needed to identify underlying structural parameters could be described by the technique. It is now a very prevalent model of time series modeling. The vector autoregression (VAR) model extends the idea of univariate autoregression to k time-series regressions, where the lagged values of all k series appear as regressors. In general, a vector of time series variables is regressed on lagged vectors of these variables. VAR (p) model of two variables X_t and Y_t is given by the equation (5) and (6)

$$Y_t = \beta_{y0} + \beta_{yy1} y_{t-1} + \beta_{yyp} y_{t-p} + \cdots + \beta_{yx1} x_{t-1} + \beta_{yxp} x_{t-p} + v_t^y \quad (5)$$

$$X_t = \beta_{x0} + \beta_{xy1} y_{t-1} + \beta_{xyp} y_{t-p} + \cdots + \beta_{xx1} x_{t-1} + \beta_{xxp} x_{t-p} + v_t^x \quad (6)$$

In this equation (6), β_{xyp} represents the coefficient of y in the equation for x at lap p whereas β_{yxp} in equation (5) the coefficient of x for y at lap p. The error terms v_t^y and v_t^x in equation (5) and (6) represent the parts of Y_t and X_t that are not related to past values of the variables. These terms could be correlated with one another because there will be some movements in Y_t and X_t to be correlated for some contemporaneous causal relationship. A key feature of those equation is that if all the variables are stationary and ergodic it can produce desirable estimation.

3.3 Data Source

The input data used in this literature are images from the year 2013 to 2019 acquired from the Landsat 8 satellite and ASTER with a ground field of view ranging from 15 m to 80 m [1]. The data from the satellites were affected by cloud cover which brings some discrepancy in the detection of cell types being eliminated later from further considerations. Table 1. shows the satellite data acquired in different timeframe along with google placemarks.

Table 1. *Data Sources used in the study*

Data Type	Details	Acquisition time
LANDSAT 8 OLI Level 2	30 m * 30 m resolution	2017-05-02 2013-10-30 2015-10-04 2019-08-28
ASTER14 DEM	15m*15m Resolution	2019-12-14 2015-10-12
Road Network	Shape file	
Placemarks Google	Educational Institutions, Hospitals, Shopping Malls, Business Areas	

3.4 Pre-Processing

Landsat 8 produces 30 m * 30 m data which is converted into 15 m* 15 m resolution to enhance visualization. Later, the stretch function of ArcGIS enhances the image by changing its properties such as brightness, contrast, and gamma. Besides, the percent clip function used also removes the lowest and highest pixel values over and below 2% to make the histogram more compact.

3.5 Maximum Likelihood Event (MLE) Classification

The MLE is applied to Landsat 8 data to classify land covers. 2 types of land covers, built-up and non-built up are classified.

3.6 Post Processing

The majority filter replaces cells based on the majority of their contiguous neighboring cells. The filter removes lonely pixels after the classification. Fig. 4. shows the classification between built-up and non-built up areas from the satellite data of 2019 after the post processing.

3.7 Variable Preparation

For modeling urban expansion using SVM, the factors prepared from different spatial data collected from 2013 to 2019 are converted into variables. 9 selected factors are most widely used opted from the experiments of previous researchers [2].

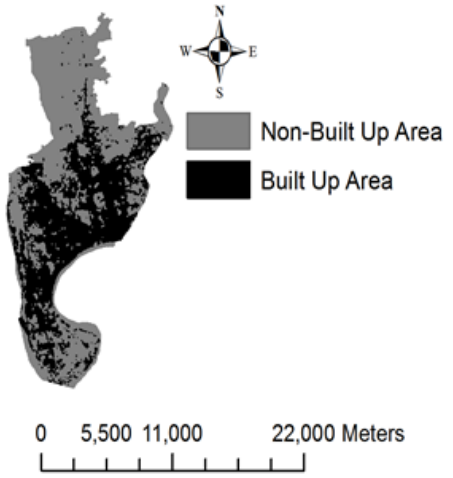

Fig. 1. Built-up and non-built up classification of satellite data of 2019.

Table 2. *shows the variables used for training the model.*

1. **Place Related Characteristics:** Four significant factors are considered in the model. Firstly, the type of cell that tells natural circumstances to the model. Secondly, a restricted variable is used to identify whether the cell belongs to any restricted area such as wildlife Conservation park, Army Cantonment or Archeological Establishments. Lastly, Aspect and slope are considered since human beings tend to live in flat areas. Therefore, less population density is seen on hills and slopes.

2. **Proximity Related Characteristics:** Proximity means the Euclidian distance of a cell from various influential socio-economic sites. The distance from the business sites, nearest hospitals, schools, shopping malls, transportation network infrastructures (roads/highways) are considered significant influencing Urbanization since human habitation expedites where those facilities become adequate. Moreover, the proximity to the nearest urban areas is also considered vital since less monetary costs are required to connect with various urban amenities such as water, sewerage, and electricity. Because of mobility and low transportation costs, new residential areas adjacent to roads and railroads emerge.

Table 2. *The variables for training the model.*

Variable Name	Description	Datatype
Dependent		
Y	1 and 0 show the presence of built up growth	Dichotomous
Independent	Place Related Variables	
TOC	Type of cell	Dichotomous

Slope	Slope	Continuous
Aspect	Aspect	Continuous
Restricted	Restricted	Dichotomous
	Proximity Related Variables	
Hdist	Distance from the nearest medical	Continuous
Sdist	Distance from the nearest educational institution	Continuous
Mdist	Distance from the nearest shopping mall	Continuous
Bdist	Distance from the nearest business area	Continuous
Rdist	Distance from road network	Continuous

3.8 Anomaly Detection and Remove

SVM model could perform inefficiently if outliers are present in the data. Therefore, the Z-score algorithm is applied to data [4]. It eliminates those data which distance is more than 3 from the mean. Moreover, the cells with cloud covers are also removed before training the model.

4 RESULT & DISCUSSION

In this literature, the SVM model is developed using the value of 10 for the C parameter known as the penalty parameter and Radial Bias function. Since the choice of the C parameter often affects classification, a combination of different values is used to assess the performance of the model. If the C parameter is large, it can cause overfitting which trains all points for the model. Therefore, it can lead to a decrease in accuracy when testing data is used. On the other hand, when the C parameter is small, SVM makes a large decision boundary where some points could be misclassified. But it increases accuracy. A 5-fold cross-validation is used so that all sample points can be used for the training and testing of the model. To find out the best model, it is developed allowing different kernel functions by changing the parameter C. For this reason, a combination of the C values of 0.1,1,10,100 and 1000 with various configurations of different kernel functions including the linear kernel, the RBF kernel with γ values of 0.01, the polynomial kernel with q values of 8 is used and tested to select the best model. Fig. 5. shows the accuracy, precision, sensitivity and specificity when C parameter varies. As illustrated in the Fig. 5, the training accuracy, precision, specificity, and sensitivity of RBF are considerably higher than those of the rest of the models. Even though the performance of the linear kernel function is quite equal with the RBF kernel,

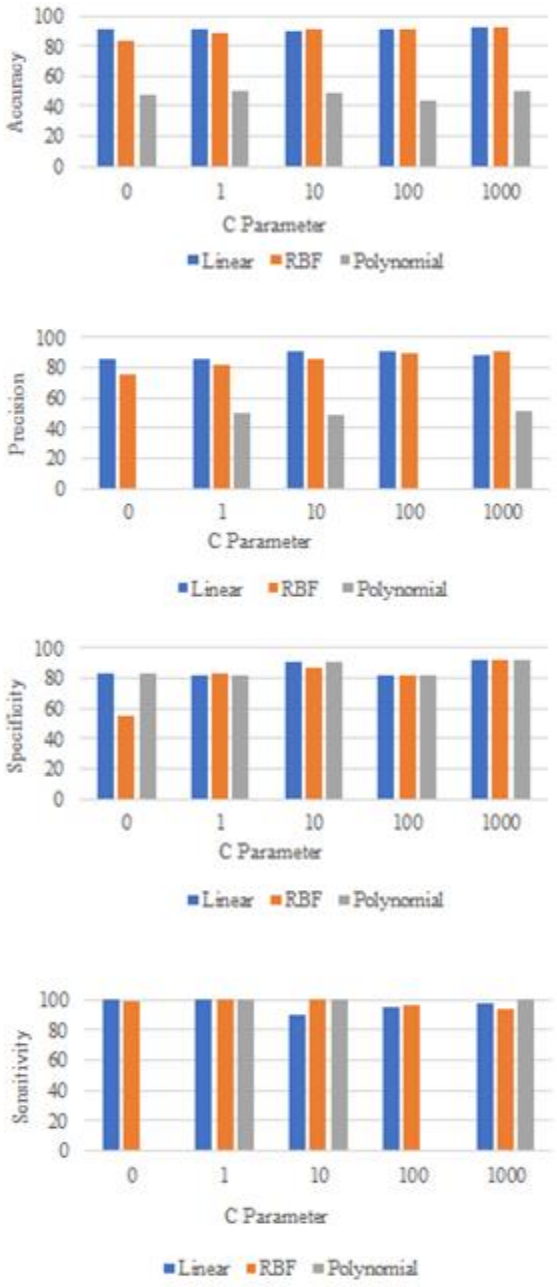

Fig. 1. Accuracy, precision, specificity and sensitivity

the RBF kernel is chosen for the model because of its capability of maximum positive prediction which increases the value of sensitivity. On the other hand, the polynomial based model fails to perform better also in the precision measurement.

Table 3. *Granger Causality Test matrix*

	Sdist_x	Hdist_x	Mdist_x	Rdist_x	Bdist_x	Slope_x	Aspect_x
Sdist_y	0.003	0.003	0.002	0.001	1.000	1.542	1.768
Hdist_y	0.000	0.0435	0.234	0.000	0.897	1.546	1.899
Mdist_y	0.0345	0.234	0.000	1.000	0.000	1.897	1.987
Rdist_y	0.3244	0.321	1.987	0.0003	0.000	1.432	1.2134

Bdist_y	0.000	1.943	0.000	0.0002	0.467	1.000	1.654
Slope_y	0.000	0.456	0.876	1.213	0.987	0.324	1.000
Aspect_y	1.2134	1.245	1.765	0.987	0.321	1.421	0.213

Besides, Vector Auto Regression (VAR), a multivariate time series forecasting algorithm, is also used to forecast the growth of variables in 2021. This model forecasts growth of those variables using the data labeled as built-up by the SVM. This algorithm is inevitable when two or more time series influence each other. In Vector Auto Regression, a time series can be affected by the past time series of another series. Therefore, it is significant to investigate whether a time series is affected by any other variables. Grangers causality, a null hypothesis tests past values of a time series X does not cause another time series Y [3],[5]. In this study, p-value 0.05 is used as a cut-off that means if the P-value is less than the cut-off, the null hypothesis gets rejected [6]. Table 3. shows a grangers causality matrix showing the p-values between the variables with x (Sdist_x, Hdist_x, Mdist_x, Rdist_x, Bdist_x, Slope_x, Aspect_x) and the variables with y (Sdist_y, Hdist_y, Mdist_y, Rdist_y, Bdist_y, Slope_y, Aspect_y). According to this table 3, a p-value less than 0.05 means the respective variable with x causes the respective variable with y. "Sdist_x", "Mdist_x","Hdist_x" and "Rdist_x" reject the null hypothesis and cause the time series calculation of "Sdist_y" since all of the variables are less than 0.05. This dependencies of a time series with the past values of other time series have made the use of VAR justifiable.

***Table 4**. AIC, BIC and HQIC for different lags*

Lag	1	2	3	4
AIC	-15.5880	-18.6212	-13.6212	-21.6264
BIC	-15.3960	-15.2687	-15.1080	-14.9520
HQIC	-15.5138	-15.4850	-15.4230	-15.3659

Different combinations of lag values from 1 to 4 are used to find out the values of Akaike information criterion (AIC), Hannan-Quinn information criterion (HQIC) and Bayesian information criterion (BIC) to decide the lag value for the VAR model. The model uses lag value 4 since it gives the minimum AIC. Table 4. shows the value of AIC, BIC, HQIC for different lags. Since time series prediction requires the data should be stationary, the VAR model ensures stationarity by first-order differencing of all-time series. A stationary time series has mean and variance which does not change over time. Moreover, it does not have any trend and seasonality effects on the data. After the data becomes stationary, the VAR model is developed with a lag value of 4. The built-up area data are split into 2 parts for training and testing. The data from 2013 to 2017 are kept for the training. The model predicts possible growth of built-up areas 2 steps ahead for the year of 2018 and 2019. The forecasted values of 2019 are compared with the expected value of 2019 to get the Root Mean Square Error (RMSE) of all the variables to determine the accuracy of the model. Table 5. is showing the RMSE of all the variables.

***Table 5**. RMSE of the variables*

Variables	RMSE	Variable	RMSE
Sdist	15.756	Hdist	12.323
Mdist	0.563	Slope	8.326
Rdist	17.673	Aspect	11.256
Bdist	21.315	Elevation	6.324

After the accuracy assessment, the data from 2013-2019 are fit altogether with the model to get the forecast of 2 steps ahead for 2020 and 2021. The new variable values of the year 2020 and 2021 are then used to extract samples that matched with the forecasted values. Table 6. shows the coefficients in different lags for the calculation of "Sdist" variable.

***Table 6**. Coefficients of the time series Sdist*

Lag	Variables	Sdist	
	Const= 0.002778	Coefficient	Probability
L1	Sdist	-0.254087	0.820
	Hdist	-0.021931	0.000
	Mdist	0.106759	0.505
	Rdist	-0.045290	0.037
	Slope	0.044280	0.612
L2	Sdist	-0.264502	0.782
	Hdist	0.003477	0.000
	Mdist	0.094875	0.918
	Rdist	0.005605	0.065
	Slope	-0.065243	0.950
L3	Sdist	-0.174953	0.684
	Hdist	0.015007	0.000
	Mdist	-0.027050	0.657
	Rdist	-0.005987	0.602
	Slope	0.089110	0.947
L4	Sdist	-0.103095	0.578
	Hdist	0.041083	0.015
	Mdist	0.039911	0.208
	Rdist	-0.023209	0.435
	Slope	0.091623	0.795

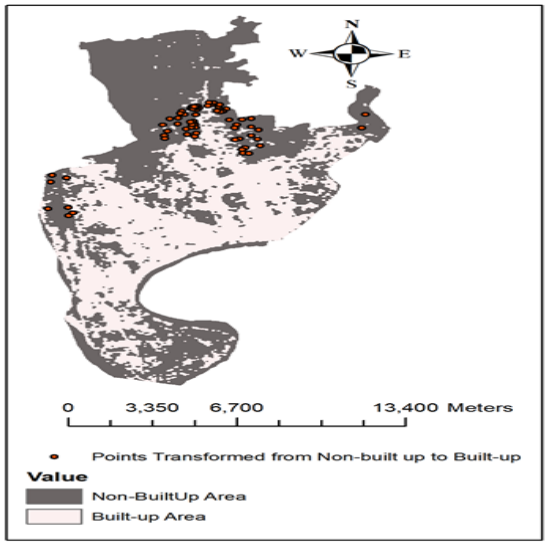

Fig. 1. Points transformed from non-built up to built-up areas in 2021

Table 7. is showing the no of affected samples, no of the current built-up area samples and no of the non-built up area

samples. It shows, 213 samples are affected by the forecast. Among those samples, 170 samples have already been classified as built-up by the SVM. However, the remaining 43 samples which are now labeled as the non-built up area could likely to be turned into the built-up area in between the year of 2020 and 2021. Fig. 6. shows locations of those transformed point on the map of year 2019.

Table 7. *The no of samples affected by the VAR model forecast, no of the current built-up area samples and no of the non-built up area samples*

Total no of samples	The no of the built-up area sample	The no of the non-built up area sample which will be transformed into the built-up area.
213	170	43

5 CONCLUSION

In this study, the capability of the SVM and VAR is explored for modeling and analyzing urban expansion and investigating the driving factors as well as their level of significance in the urban expansion process of Chittagong city corporation throughout for 2013-2019. Various functions and configurations are used together to get the most capable model that can classify urban built up and non-built up cells efficiently. The modeling process involves the examination of various variables, the investigation of SVM parameterization and Kernel regulation. The result shows the SVM model is capable of both positive and negative predictions with the highest accuracy. Similarly, the VAR model, a multivariate time series prediction is developed in the study to forecast the growth of the variables for the built-up areas in 2021. These forecasted values are later used to identify the affected samples that could turn into the built-up area. The SVM plays an inevitable role to extract possible causes and examine the impacts of urban expansion on habitants, environmental pollutions, wildlife disturbance, and deforestation. Besides the use of VAR alongside can forecast how much effects will be caused in the imminent future. The models altogether could become a useful tool for urban planners, geographers, and environmental policy makers to identify the interaction between nature and artificial environment.

6 ACKNOWLEDGMENT

The authors wish to thank earthexplorer.usgs.gov, Goggle API for the satellite data. They also wish to thank Md. Rakibul Sohel for proofreading.

7 REFERENCES

[1]. Earthexplorer.usgs.gov, retreived from https://earthexplorer.usgs.gov/, 2019

[2]. Firoozeh K., Selima S., Ali Shirzadi B., Shan Suthaharan," An enhanced support vector machine model for urban expansion prediction", ELSEVIER, Computers, Environment and Urban Systems 75, 61-75, pp. 7-8, (2019).

[3]. Petros D.,"Using multivariate cross correlations, Granger causality and graphical models to quantify spatiotemporal synchronization and causality between pest populations", Springer, pp. 6-8, August 2016.

[4]. Alexander E. Curtis, MBChB, BSc1 , Tanya A. Smith, MBChB, BSc1 , Bulat A. Ziganshin, MD1,2, John A. Elefteriades, MD1, "The Mystery of the Z-Score", AORTA, August 2016, Volume 4, Issue 4:124-130, pp. 125-127, , June 2016.

[5]. AC Jordaan, JH Eita," Export and economic growth in Namibia: a Granger causality analysis", South African Journal of Economics, pp. 3-4, 2007.

[6]. blog.minitab.com, retrieved from https://blog.minitab.com/blog/understanding-statistics/what-can-you-say-when-your-p-value-is-greater-than-005, , (2019).