Using Multivariate Adaptive Regression Spline and Artificial Neural Network to Simulate Urbanization in Mumbai, India

M. Ahmadlou [a] , M. R. Delavar [b*], A. Tayyebi [c] and H. shafizadeh- Moghadam[d]

[a] GIS Dept., School of Surveying and Geospatial Eng., College of Eng., University of Tehran, Tehran, Iran, Email: m_ahmadlou@ut.ac.ir
[b] Center of Excellence in Geomatic Eng. in Disaster Management, School of Surveying and Geospatial Eng., College of Engineering, University of Tehran, Tehran, Iran, Email: mdelavar@ut.ac.ir
[c] University of California-Riverside, Center for Conservation Biology, Riverside, CA, USA, Email: amin.tayyebi@gmail.com.
[d] Tarbiat Modares University, Department of GIS & RS, Tehran, Iran, Email: hshafeezadeh@gmail.com

a) **KEY WORDS:** Land Use Change, Data Mining, Multivariate Adaptive Regression Spline, Artificial Neural Network, Receiver Operating Characteristic

ABSTRACT:

Land use change (LUC) models used for modelling urban growth are different in structure and performance. Local models divide the data into separate subsets and fit distinct models on each of the subsets. Non-parametric models are data driven and usually do not have a fixed model structure or model structure is unknown before the modelling process. On the other hand, global models perform modelling using all the available data. In addition, parametric models have a fixed structure before the modelling process and they are model driven. Since few studies have compared local non-parametric models with global parametric models, this study compares a local non-parametric model called multivariate adaptive regression spline (MARS), and a global parametric model called artificial neural network (ANN) to simulate urbanization in Mumbai, India. Both models determine the relationship between a dependent variable and multiple independent variables. We used receiver operating characteristic (ROC) to compare the power of the both models for simulating urbanization. Landsat images of 1991 (TM) and 2010 (ETM^+) were used for modelling the urbanization process. The drivers considered for urbanization in this area were distance to urban areas, urban density, distance to roads, distance to water, distance to forest, distance to railway, distance to central business district, number of agricultural cells in a 7 by 7 neighbourhoods, and slope in 1991. The results showed that the area under the ROC curve for MARS and ANN was 94.77% and 95.36%, respectively. Thus, ANN performed slightly better than MARS to simulate urban areas in Mumbai, India.

1. INTRODUCTION

Unprecedented urban growth, one of the most common land use change (LUC) forms, especially in developing countries has caused the control of metropolitan areas to become out of the hands of urban policy makers and planners (Tayyebi et al., 2011; Pijanowski et al., 2010). For example, urbanization has led to the expansion of urban areas from previously open areas that were originally natural areas and agricultural lands (Pijanowski et al., 2014). As a result, disturbance of agriculture and forest areas affected food security of human populations and reduced the richness of biodiversity, respectively (Tayyebi and Pijanowski, 2014). At a global scale, extensive conversion from vegetation to agriculture mainly occurred on shrub land with lack of forage production during the 1970s and into the mid-1980s (Armesto et al., 2009). However, after the mid-1980s, agricultural growth occurred with some degree of intensification in those areas more suitable for agriculture. Some recovery of forests from shrub lands and abandoned agricultural land has occurred recently (Kolmogoro, 1956).

In the recent years, due to the effects LUC can have on the environment and human life, LUC modeling is considered as a significant issue. Many factors involved in LUCs including demographic (e.g., population growth), economic (e.g., gross domestic product), bio-physical parameters (e.g., elevation and soil), institutional issues (e.g., policies) and cultural affairs. LUC causes changes in climate (Watson et al., 2000), economy (Long et al., 2007), food security (Godfray et al., 2010), water cycle (Tayyebi et al., 2015), which are threats to human life and well-being. Therefore, it is essential to examine this phenomenon. To study LUC, information about LUC drivers is needed. These drivers operate at temporal and spatial scales and occur in a non-linear manner (Veldkamp and Lambin, 2001). So, it requires precise and advanced techniques to model LUC (Tayyebi et al., 2014). Given the complexities of LUC, using data mining techniques to understand the hidden patterns in land use data can help to understand this process better. In a general classification, data mining models can be divided into two main groups: 1) global parametric models, and 2) local non-parametric models. Global parametric models use all the available data for LUC modeling (Theobald and Hobbs, 1998), also, model structure is fixed before modeling. Artificial neural

* mdelavar@ut.ac.ir

network (ANN) widely used by LUC modelers is a global parametric model. In contrast, local non-parametric models divide data into subsets and then apply the modeling procedure to each subset. Also, model structure is not fixed before modeling. Multivariate adaptive regression spline (MARS) is a local non-parametric model introduced in the literature as method provides the best fitness for a given dataset. Although global parametric models are used more than local non-parametric models, there are few studies that compared these two models together. So, in this study, we compared MARS with ANN. Relative Operating Characteristics (ROC) was used to compare the accuracy of the two models. In this study, Landsat images of 1991 (TM) and 2010 (ETM^+) were used for modelling urbanization. These images were obtained from the United States Geological Survey (USGS) portal.

2. METHODS

2.1 MARS

MARS (Friedman, 1991) is a non-parametric model that divides data into various partitions and formulates the relationship between independent and dependent spatial drivers (Tayyebi and Pijanowski, 2014). This relationship was established using piecewise polynomial functions called basis functions (Friedman, 1991). In contrast to other non-linear models where fit only one set of coefficients to the data, MARS fits separate piecewise polynomial functions to each region and creates a separate set of coefficients (Tayyebi et al., 2014). Furthermore, complex non-linear interactions between spatial drivers of LUC can also be specified. The general model of MARS is in the form of Eq. (1) (Friedman, 1991):

$$\hat{Y} = \hat{f}(\mathrm{X}) = \sum_{m=1}^{M} \alpha_m B_m(\mathrm{X}) + e \quad (1)$$

where M= the number of sub-regions
m = the number of spatial drivers of LUC
e = the error term
α= the basis function coefficients
X= the independent variables
Y= the land use classes

B is the basis functions which can be represented as (Friedman, 1991):

$$B_m(\mathrm{X}) = \prod_{i=1}^{N_m} [S_{i,m}(X_{v(\mathrm{i,m})} - t_{i,m})]_+^q \quad (2)$$

where N= the interaction order of the mth basis function
$S_{i,m} = \pm 1$
$X_{v(i,m)}$ = the vth variable where $1 \le v(i,m) \le k$
k= the total number of spatial drivers
$t_{i,m}$ = a knot location of the spatial drivers
q = the power of the basis function

N can be specified by a user given to prior knowledge about the application. When q = 1, simple linear splines are selected. The subscript '+' is according to the following phrase (Friedman, 1991):

$$[S_{i,m}(X_{v(i,m)} - t_{i,m})]_+^q = \begin{cases} [S_{i,m}(X_{v(i,m)} - t_{i,m})]_+ & S_{i,m}(X_{v(i,m)} - t_{i,m}) \ge 0 \\ 0 & otherwise \end{cases} \quad (3)$$

The objective of MARS is to minimize the sum of the square errors to regulate the basis function coefficients. MARS may over-fit data in the training phase by adding dispensable basis functions to the model. To avoid this over-fitting, the basis functions with the least contributions are eliminated by using generalized cross validation (GCV) (Friedman and Silverman, 1989) as such:

$$GCV = \frac{(1/n)\sum_{i=1}^{n}[y_i - f(X_i)]^2}{[1-(\mathrm{C(m)}/n)]^2} \quad (4)$$

where n = number of total observations
y = response variable
f = estimated function by MARS

The aim of MARS is to minimize the GCV and the best model is the one with the lowest GCV. C(M) is as follow (Friedman, 1991):

$$C(\mathrm{M}) = \mathrm{C} \times \mathrm{d} \quad (5)$$

where d = the cost for each basis function
M = total number of basis functions

2.2 Artificial Neural Network

Artificial neural network is one of the most common global parametric models applied to model LUC (Pijanowski et al., 2002, 2010 and 2014). A multi-layer perceptron (MLP) is designed to identify an unknown relation between spatial drivers of the LUC and land use classes. The designed MLP consists of three layers (an input and an output layer with one hidden layer). Neural network used in the model has 10 input nodes, 21 nodes in the hidden layer and 1 node for output layer(Kolmogoro, 1956). ANN used delta method for adjusting error between nodes.

2.3 Relative Operating Characteristic

ROC eliminates the limitation of defining a unique threshold for the problem using different thresholds. After defining thresholds, these values are applied to the suitability map (map shows the suitability of change for each cell which varies from low to high). The values greater than these thresholds in the suitability map (a map showing the membership of each cell to either change or no-change) are set to 1 (meaning LUC in the desired pixel) and the other values are set to zero (meaning non-LUC in the desired pixel). The result is compared with the real LUC map. Then, the 2×2 contingency tables are calculated based on the Table 2 for each threshold. In this table, True negative (TN) indicates cells which are forecasted as non-change and are actually non-change in the reference map. False positive (FP) indicates cells which are forecasted as non-change but are actually change in the reference map. False negative (FN) indicates cells which are forecasted as change but are actually non-change in the reference map. Finally, True positive

(TP) indicates cells which are forecasted as change and are actually change in the reference map. After generating the contingency tables, we calculated X_t and Y_t for different thresholds to plot the ROC curve according to Eq. (6) the following relations:

$$X_t = \frac{FN}{FN + TN}, Y_t = \frac{TP}{TP + FP} \quad (6)$$

The Area Under ROC Curve (AURC) is obtained as:

$$AURC = \sum_{t=1}^{T-1} [Y_{t+1} + Y_t][X_{t+1} - X_t]/2 \quad (7)$$

where T= shows total number of the thresholds
t = the threshold values

2.2 Study Area

Mumbai is the capital of Maharashtra State located in the western part of India and has the 7th highest population density in the world. Both images were classified into urban, bodies of water, wetland, forest and agricultural classes using maximum likelihood classification method based on Anderson level 1 (Anderson et al. 1976; Shafizadeh-Moghadam and Helbich, 2013; Figure 1). Accuracy of the classified images was examined using Kappa index, based on 250 randomly selected points. The accuracy of classified images for 1991 and 2010 was 84% and

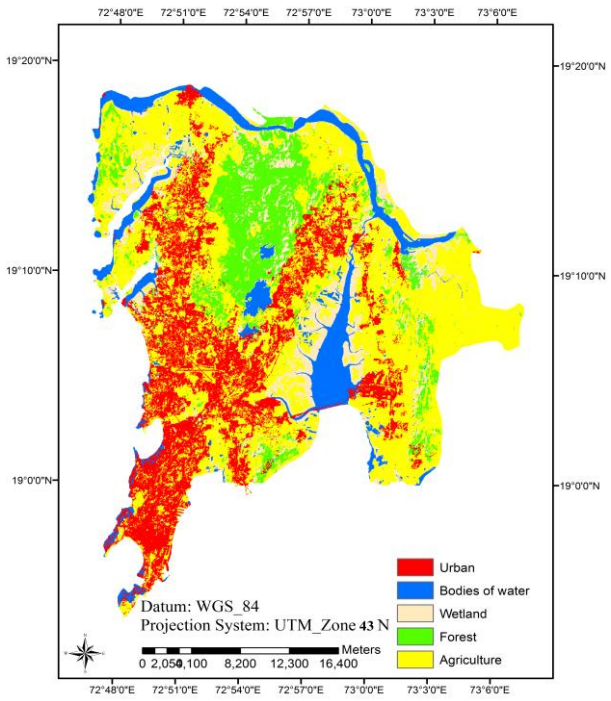


Figure 1. Study area in 1990

	Predictors in 1990
1	Elevation
2	Distance to built areas
3	Density of built areas
4	Distance to road
5	Distance to bodies of water
6	Distance to forest
7	Distance to railway
8	Distance to central business district
9	Number of agricultural cells in a 7 by 7 neighbourhood
10	Slope

Table 1. List and the date of spatial predictors in study area

86%, respectively. 60% of the entire data was used for training run and rest of the data was used to simulate the urban pattern at 2010. Factors considered as drivers of change between 1990 and 2010 are listed in Table 1. All of the land use drivers which were considered as inputs to the MARS model were prepared in GIS environment.

3. RESULTS AND DISCUSSION

3.1 MARS

Table 3 presents inconstant effect of each variable in urban gain. For example, the CBD effect is negative for distances less than 19,988.42 m and positive for the other intervals. The likelihood of urban gain increased sharply for the distances between 0 and 19,988.42 m. Also, the likelihood of urban gain declined slowly (smaller coefficient) for the distances between 19,988.42 m and 35000 m (Figure 2). Similarly, Figure 2 shows the effect of other LUC drivers in urban gain. In addition, the coefficient of each LUC drivers is given in Table 3**.** MARS found one knot (around 19,988.42 m) or two sub- regions (Figure 2) for Distance to central business district (Table1). Stratified random sampling used to extract 60% of data for training and 40% for testing.

3.2 ANN

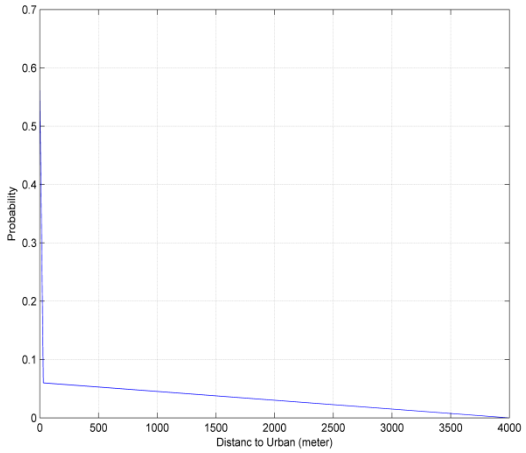
We obtained 0.04241 for mean squared error in 500 cycles. Figure 3 shows training error for ANN.

3.3 ROC

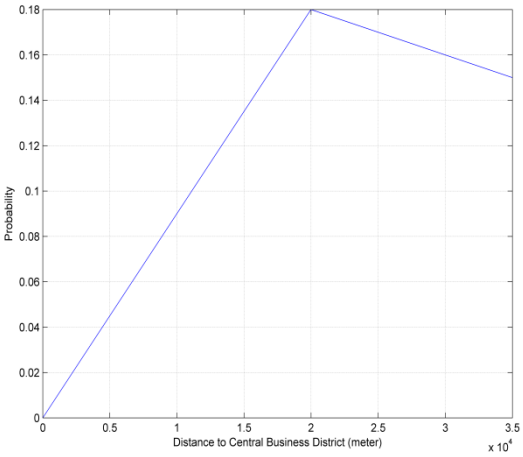
Performance accuracy of the MARS and ANN models were evaluated using ROC. The area under the ROC curve was 94.77% and 95.36% for MARS and ANN models, respectively (Figure 4).

Simulated map	Reference map		
	Change	Non-change	Reference Total
Change	True Positive (TP)	False Negative (FN)	TP + FN
Non-change	False Positive (FP)	True Negative (TN)	FP + TN
Diagnosis Total	P = TP + FP	Q = FN + TN	P + Q

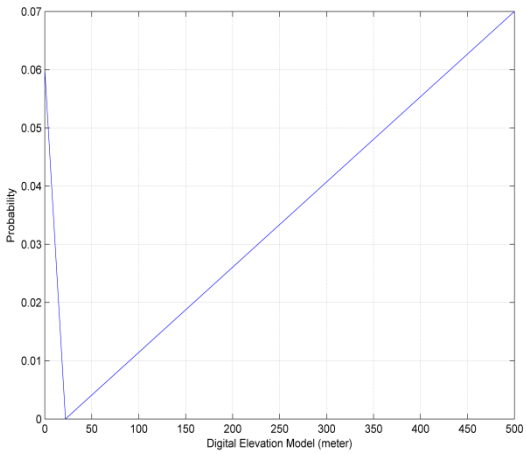
Table 2. Comparing simulated map and reference man using contingency tables



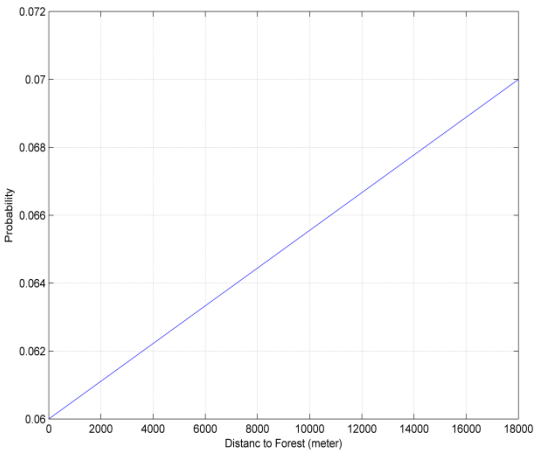
a) Distance to urban



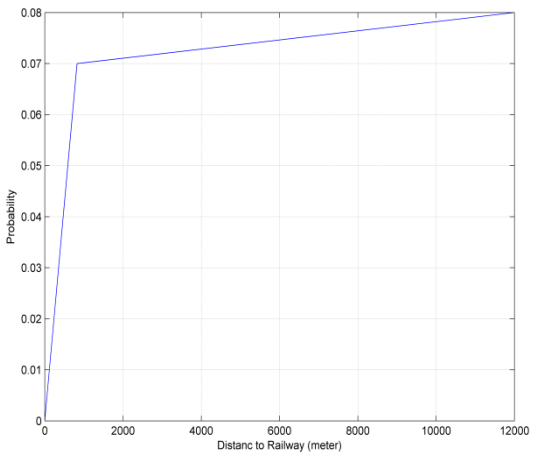
b) Distance to central business district



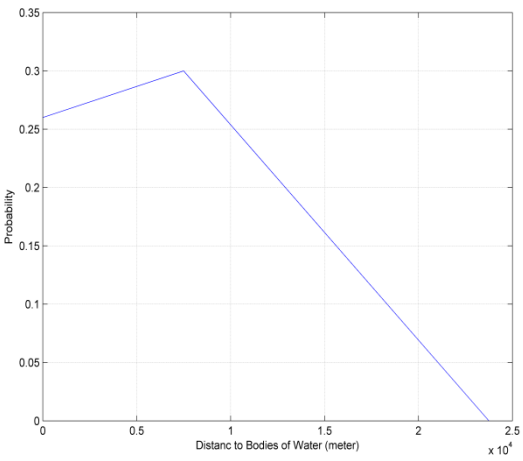
c) Digital elevation model



d) Distance to forest



e) Distance to railway



f) Distance to water bodies

Figure 2. Basic functions of MARS for significant drivers

BFs	Coefficients	Variable	Sign	Knot (meter)
0	0.1274			
1	-0.0026	Urban	+	30.3440
2	0.0139	Urban	-	30.3440
3	-0.0143	Durban	+	27.0000
5	-0.0000	CBD	-	19,988.42
6	-0.0000	CBD	+	19,988.42
7	-0.0000	Water	-	8,109.564
8	-0.0000	Water	+	8,109.564
9	0.0026	Urban	-	67.8512
11	0.0000	Railway	+	824.3293
12	-0.0001	Railway	-	824.3293
13	0.0001	Elevation	+	22.0000
14	0.0027	Elevation	-	22.0000
15	0.0000	Forest	+	0.0000

Table 3. Coefficients, variables and knots of MARS in the study area

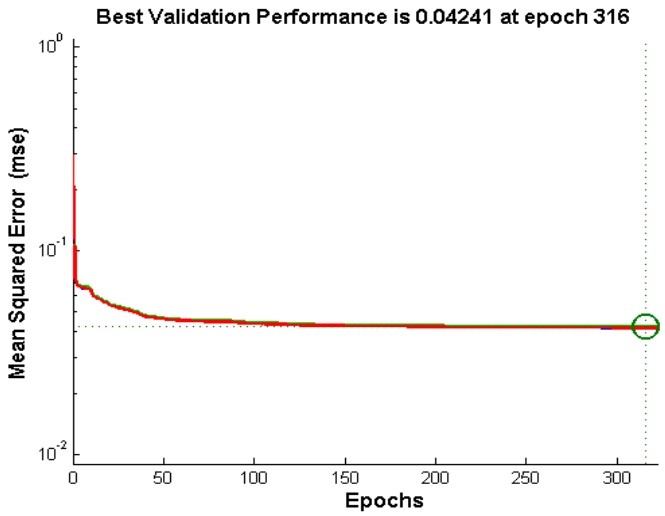
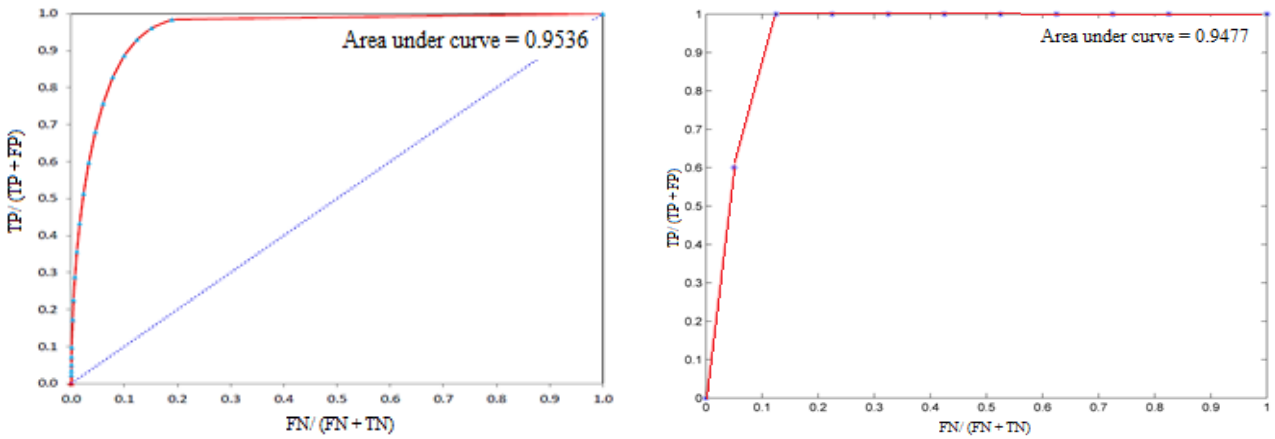


Figure 3. Training run for ANN



a) ROC curve of ANN b) ROC curve of MARS

Figure 4. ROC curves and area under curves (AUC) of the model for a) ANN and b) MARS models.

3. CONCLUSION

This study presented and compared two models, MARS and ANN, to simulate urbanization in Mumbai city of India. The considered drivers for urbanization in this area were distance to urban areas, urban density, distance to roads, distance to bodies of water, distance to forest, distance to railway, distance to central business district, number of agricultural cells in a 7 by 7 neighbourhood and slope in 1991. The results showed that the area under the ROC curve for MARS and ANN were 94.77% and 95.36%, respectively. Thus, ANN performed slightly better than MARS to simulate urban gain.

4. REFERENCES

Armesto, J.J., Smith-Ramírez, C., Carmona, M.R., Celis-Diez, J.L., Díaz, I.A., Gaxiola, A., Gutiérrez, A.G., Núnez-Avila, M.C., Pérez, C.A., Rozzi, R., 2009. Old-growth temperate rainforests of South America: conservation, plant–animal interactions, and baseline biogeochemical processes, in: Old-Growth Forests. Springer, pp. 367–390.

Friedman, J.H., 1991. Multivariate adaptive regression splines. Ann. Stat. 1–67.

Friedman, J.H., Silverman, B.W., 1989. Flexible parsimonious smoothing and additive modeling. Technometrics 31, 3–21.

Godfray, H.C.J., Beddington, J.R., Crute, I.R., Haddad, L., Lawrence, D., Muir, J.F., Pretty, J., Robinson, S., Thomas, S.M., Toulmin, C., 2010. Food security: the challenge of feeding 9 billion people. science 327, 812–818.

Kolmogoro, A.N., 1956. On the representation of continuous functions of several variables as superpositions of functions of smaller number of variables, in: Soviet. Math. Dokl. pp. 179–182.

Long, H., Heilig, G.K., Li, X., Zhang, M., 2007. Socio-economic development and land-use change: Analysis of rural housing land transition in the Transect of the Yangtse River, China. Land Use Policy 24, 141–153.

Shafizadeh-Moghadam, H., & Helbich, M. (2013). Spatiotemporal urbanization processes in the megacity of Mumbai, India: a Markov chains-cellular automata urban growth model. Applied Geography, 40, 140-149.

Pijanowski, B.C., Brown, D.G., Shellito, B.A., Manik, G.A., 2002. Using neural networks and GIS to forecast land use changes: a land transformation model. Comput. Environ. Urban Syst. 26, 553–575.

Pijanowski, B. C., Tayyebi, A., Delavar, M. R., & Yazdanpanah, M. J. (2010). Urban expansion simulation using geospatial information system and artificial neural networks.

Pijanowski, B. C., Tayyebi, A., Doucette, J., Pekin, B. K., Braun, D., & Plourde, J. (2014). A big data urban growth simulation at a national scale: Configuring the GIS and neural network based Land Transformation Model to run in a High Performance Computing (HPC) environment. Environmental Modelling & Software, 51, 250-268.

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 and Urban Planning, 100(1), 35-44.

Tayyebi, A., Pijanowski, B.C., Linderman, M., Gratton, C., 2014. Comparing three global parametric and local non-parametric models to simulate land use change in diverse areas of the world. Environ. Model. Softw. 59, 202–221.

Tayyebi, A., & Pijanowski, B. C. (2014). Modeling multiple land use changes using ANN, CART and MARS: Comparing tradeoffs in goodness of fit and explanatory power of data mining tools. International Journal of Applied Earth Observation and Geoinformation, 28, 102-116.

Tayyebi, A., Pijanowski, B.C., Pekin, B.K., 2015. Land use legacies of the Ohio River Basin: Using a spatially explicit land use change model to assess past and future impacts on aquatic resources. Appl. Geogr. 57, 100–111.

Theobald, D.M., Hobbs, N.T., 1998. Forecasting rural land-use change: a comparison of regression-and spatial transition-based models. Geogr. Environ. Model. 2, 65–82.

Veldkamp, A., Lambin, E.F., 2001. Predicting land-use change. Agric. Ecosyst. Environ. 85, 1–6.

Watson, R.T., Noble, I.R., Bolin, B., Ravindranath, N.H., Verardo, D.J., Dokken, D.J., others, 2000. Land use, land-use change and forestry: a special report of the Intergovernmental Panel on Climate Change. Cambridge University Press.