Kalpa Publications in Computing
Volume 13, 2019, Pages 96–104
Proceedings of the 1st International Conference on Geospatial Information Sciences

Automatic Modelling of Land Use Suitability Using Deep Feedforward Networks in Leon - Silao, Guanajuato Region

Rodrigo Lopez-Farias[1,a], Juan A. Pichardo-Corpus[1,b], and Raul A. Aguilar-Vilchis[2]

[1] CONACYT - CentroGeo, Centro de Investigación en Ciencias de Información Geoespacial, Contoy 137, Col. Lomas de Padierna 1420,
Delegación Tlalpan, CDMX, México
{[a]rlopez,[b]jpichardo}@centrogeo.edu.mx
[2] Universidad Autónoma de Querétaro, Facultad de Ciencias Políticas y Sociales,
Cerro de las Campanas s/n 76010, Querétaro, Querétaro, México
raula4820@gmail.com

Abstract

Land use change is a global phenomenon that impacts directly to the urban growth and it should be addressed from different disciplines to minimize the potential negative effects of urbanization predicting the spatial urban growth. The urban growth dynamics might be very complicated and difficult to model, nevertheless it is necessary to understand the causes of the dynamics and the dynamics itself to build precise computational models that help to detect problems generated by urban land use change. For that reason, we propose the study of the land use suitability sub-model used by several models to make land use spatial predictions. This sub-model is implemented as a logistic regression based on linear correlations. The problem is that this model is limited to capture a variety of nonlinear relations among variables for prediction and classification purposes. We propose to use an alternative based on Deep Feedforward Networks able to deal with this problem.

In Mexico, the urban growth will increase considerably the number of cities during the next decade where the Mexican population will be concentrated. That means that the generation and study of existing spatio-temporal computational frameworks for studying the Mexican urban growth is very relevant. Therefore we present an initial contribution comparing Deep Feedforward Networks with a Multi-Level Linear Logistic Regression as land suitability models applied to Mexican land use classification. We show that basic deep feedforward models outperform in allocation accuracy to linear logistic regression, and also minimizes the parameters tuned by trial and error.

1 Introduction

Land use change is a global phenomenon that impacts directly to the urban growth. This fact makes to face key challenges in relation to sustainable urban planning. The study of this phenomenon, requires the interaction of several disciplines (e.g., sociology, geography, physics, mathematics, computer science) in order to understand the different causes and effects of the

O. S. Siordia, J.L. Silván Cárdenas, A. Molina-Villegas, G. Hernandez, P. Lopez-Ramirez, R. Tapia-McClung, K. González Zuccolotto and M. Chirinos Colunga (eds.), iGISc 2019 (Kalpa Publications in Computing, vol. 13), pp. 96–104

urban growth dynamics that produces land use changes from non urbanized (undeveloped) to urbanized (or developed) land.

The causes of land use change are diverse and several facts have been identified for instance: socio-economic, environmental, infrastructure, among others [11]. The urban growth dynamics is difficult to model and simulate due its complexity and stochasticity. Nevertheless it is inherently necessary to build computational models that helps to anticipate problems generated by urban land use change and to preserve control in urban planning as much as possible. The detection of potential undesirable growth dynamics that induces unsustainable conditions should be validated with a statistical and mathematical computer model. For example in Mexico, is expected to have an increment in the number of cities, from 384 registered in 2017 to 961 in 2030, where will live more than the 80% of the Mexican population [12]. The lack of urban planning would accelerate the increment of average time in urban mobility and the appearance of emergent and irregular urban slums with poor living conditions [2].

In this work we study a component that is included in several urban growth models (UGM) called Land Use Suitability. This component produces useful estimations with regards to the probability of land use change. The smallest spatial unit to compute in a raster map is the cell that is a squared area of certain dimension that depends on the image resolution. The land use change probabilities are associated directly with each cell and, in turn, each cell with the local characteristics of the landscape represented with a vector of variables. This vector may include information such as distance to water, distance to urban area, slope among others.

One problem is to define the relation between the urban land change (dependent) variable and the vector of (independent) variables of the local landscape characteristics. One popular approach is to assume that the land use suitability is defined mainly by linear relations among variables. For that reason linear models such as the Linear Multi-Level Logistic Regression (MLLR) are used in several models as Land Use Suitability model [11]. This kind of models are very inflexible and limited capturing other kinds of nonlinearities found in real data. For that reason we propose the implementation of a model based on Deep Feedforward Network (DFFN), that is able to associate different variables with nonlinear relations to land use change. We also show that DFFN is easier to implement in the sense that is possible to discard parameters difficult to tune by trial and error compromising the accuracy of the model, as is the case with the Land Use Suitability model implemented in FUTure Urban-Regional Environment Simulation (FUTURES) [9]. This model requires to choose two parameters by a trial and error process, associated with the pressure and the composite suitability score functions defined in Section 4. Therefore we propose a comparison of the two approaches to build Land Use Suitability Models measuring the allocation accuracy of new urban cells.

We compare a Linear Site Suitability Model based on The Multi-Level Logistic Regression (MLLR), implemented in Futures but used by other UGMs, with two nonlinear model variations based on a Deep Feedforward Network (DFFN): one DFFN uses the development pressure variable as part of its input, and the other model discards that variable. We use real georeferenced data associated with the region of Leon - Silao Municipalities of Guanajuato state in Mexico containing more than 7 million vectors. Our experiments exhibits better allocation accuracy in general by DFFN than MLLR.

This paper is structured as follows: Section 2 describes popular urban growth models implemented and available for practitioners, with their common weakness in relation with the linear models. Section 3 describes the Leon-Silao regional data. Section 4 describes the methods and data used for the experiment. Section 5 addresses the experiment and results, and in the last Section 6 we propose the future work based on the results.

2 Related Work

The experiments in this work are motivated by the work of Pickard et. al. [11], whom presented an accuracy study of different models implemented in open source software that is accessible for practitioners. The study includes four urban growth models: SLEUTH [3], LCM [6], GeoMod [5] and FUTURES [9]. SLEUTH is a stochastic cellular automaton able to deal with spatial information associated with the slope, land use, exclusion, urban, transportation and hill shading variable, (the initial of each variable gives its name). This model presented the highest allocation accuracy but a low land use change quantity accuracy. This model is inflexible since it is difficult to include other kind of variables that could influence the dynamics.

On the other hand, the models GeoMod, LCM and FUTURES that were better in predicting the quantity of urban land use change, are more flexible with regards to input data, but less accurate in allocation. These models (except SLEUTH) have in common the implementation of a Linear Logistic Regression for the estimation of land use suitability. The quantitative accuracy and flexibility of these models, suggests that it is worth to study the cause that produces bad allocation accuracy, and to determine if modifying the land use suitability submodel the accuracy is improved.

These models also share the use of several components to perform the spatio-temporal prediction task. Mainly, two of them are the land use suitability and the urban simulation components. The first one estimates the land suitability with a Linear MLLR model required to decide stochastically the location of an initial land use change (called seed) required by the second model to perform urban growth simulations.

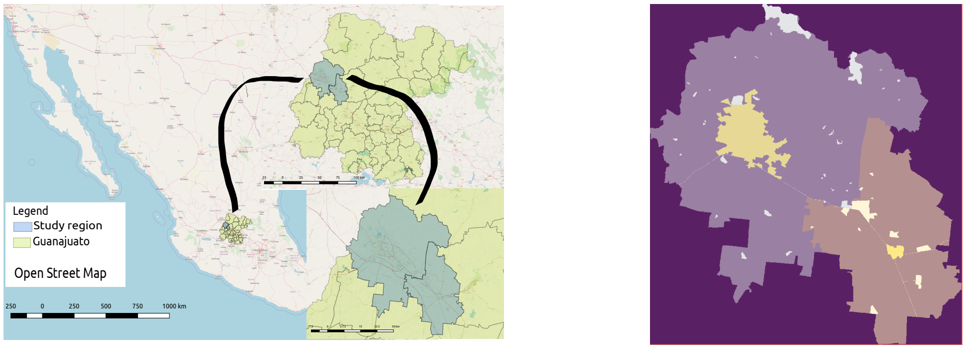
This dependency makes important to improve the accuracy of urban suitability to perform more accurate urban simulation since the land use suitability represents an initial probability of land use change for each cell. Therefore we propose to analyze the site suitability modelling and accuracy from the classification perspective using Deep Feedforward Network compared with the MLLR.

3 Data Description

The regional data of Leon-Silao municipalities are obtained from open databases hosted by the Institute of Statistics and Geographic Information (INEGI) [8]. These base maps are required to compute variables of interest that describe the land use characteristics for modelling purposes. The data is geographically constrained by the regional boundaries of Leon and Silao municipalities located in Guanajuato State, presented in Figure 1a. The resolution of the raster geo-referenced map is $30m \times 30m$ per cell.

The base raster maps used for the experiment are: municipalities boundaries, urban development at year 2000 and 2010, industrial parks, water shapes, protected natural areas, roads, rail roads and elevation. Using the base maps we compute the maps of urban pressure, distance to urban area, distance to industry, distance to water shapes, natural areas, distance to roads and rail roads, road density, travel time to city, slope and pressure. Table 1, enumerates and describes each map. The developed region taken into account is the urban land use detected from year 2000 to 2010.

These maps are computed with GrassGis Software [7], and the pressure map is computed with *devpressure* function provided by the *r.futures package* [10]. The *r.futures package* consists of eight functions: *r.futures.pga*, *r.futures.potential*(see Equation 2), *r.futures.potsurface*, *r.futures.demand*, *r.futures.devpressure*(see Equation 3) and *r.futures.category*. Each one is used to perform one part of the prediction, nevertheless for our goals, we only considered the



(a) (b)

Figure 1: Leon-Silao description: (a) Location of Leon-Silao in Mexico. (b) Excluded regions. Yellow: Development at year 2000. White: industrial park, natural areas, water shapes

Table 1: Different map layers used to define the independent variables

	Map	Required base map
1	Pressure	Urbanization in 2000
2	Distance to urban	Urban centers
3	Distance to industrial park	Industry map
4	Distance to water shapes	Water bodies map
5	Distance to natural areas	Natural areas
6	Distance to roads	Roads
7	Distance to roads	Roads
8	Distance to interchanges	Interchanges
9	Road density	Roads
10	Slope	Elevations
11	Time travel to cities	Roads + city center
12	Sub-regions	Map of municipalities

devpressure and the potential as explained below.

4 Methods

In this section we describe the two approaches for land use suitability change models: the model based on the Multi-Level Logistic Regression, implemented in several urban growth prediction models, and the Deep Feedforward Network (DFFN).

The suitability is an important measure represented by the likelihood of land use change potentially induced by a set of variables that represents the local characteristics of the landscape suitable to develop a cell. To compute the likelihood, it is important to find a model that associates as best as possible the land use change with these variables.

According to [9], these variables might be, socioeconomic, infrastructural or geographical such as slope, distance to water resources, distance to city centers, distance to urban region, road density among others. In order to find the relation between the land use change and the landscape variables, it is important to code this information. One way to do it, is coding the urban use with a binary variable for classification purposes $Y \in \{0, 1\}$, where Y is 0, to represent a non-urbanized or undeveloped cell, or 1 otherwise.

We constraint the land use change only for cells that change unidirectionally from non-urban to urban use in a time lapse determined by the time between two available urban maps. The environment of each cell i is represented by a $n-$ dimensional vector $X \in \mathbb{R}^n$ where n is the vector size representing the landscape variables. The problem of finding a good suitability model can be seen as an optimization problem,

$$\underset{\theta}{min} \sum_{i=1}^{n} (Y_i - F(X_i; \theta))^2 \tag{1}$$

, where F is a function that returns a binary variable, and Θ are the parameters that minimize the error produced by the squared sum of the differences generated by the observed and estimated output value. Next we describe briefly the two models that can be implemented as the $F(X, \Theta)$ classifier function. The Multi-Level logistic regression model and the Deep Feedforward Neural Network.

4.1 Site Suitability Model based on The Multi-Level Logistic Regression

The core of this Linear Multi-level Logistic Regression model is the regression function

$$s'_i = a_j[i] + \sum_{h=1}^{m} \beta_{j[i]h} \cdot x_{ih} + \beta_{j[i]} \cdot p'_i, \tag{2}$$

where the potential s'_i is the dependent variable of the different predictor variables, j is the level associated with the group of cells in a subregion (i.e. the municipality), h is a predictor variable, m is the number of landscape predictive variables, $a_{j[i]}$ and β_h are the parameters Θ, representing the intercept and regression coefficients, and p'_i is the development pressure. This last variable is computed with

$$p'_i = \sum_{k=1}^{n_i} \frac{State_k}{d_{ik}^{\gamma}}, \tag{3}$$

where $State_k \in \{0, 1\}$ is a binary variable representing the urban developed or undeveloped cell, d_{ik} is the distance between the k_{th} neighbor and the current cell i, and finally γ controls the neighborhood extension from to cell i. Equation 3 is used to build the Pressure map listed in Table 1.

The next equation calculate the composite suitability score s_i, which considers the distance influence

$$s_i = s'_i * d^{-\alpha}, \tag{4}$$

where s'_i, is the potential computed with Equation 2, d is the distance from the developed cell and α is a tunable parameter that makes the suitability scores decay exponentially. Finally s_i is mapped with a logistic function to produce a probability

$$p_i = \frac{e^{s_i}}{1 + e^{s_i}}. \tag{5}$$

One advantage of this model is its interpretability because each relation can be analyzed with linear correlations. In order to fit no lineal data, it uses two data manipulation techniques: The first one is the use of the exponential functions in Equation 4 and 3. These functions describe the pressure and suitability strength as a kind of gravitational function regulated by γ and α respectively.

The other numerical manipulation is the multi-level component, which makes to each level behaves like a independent sub-model that fits a relation between the input and output vectors of a municipality bounded area. This makes to have different coefficient values for each region. Although this model is simple and relatively easy to optimize due its linear structure, α and γ parameters should be selected by trial and error. The linear structure itself makes inflexible to fit a broader kind of nonlinear relations potentially presented in multi-dimensional data.

4.2 Deep Feedforward Network Model

In contrast with previous method, the Deep Feedforward Networks are computational models inspired in neuroscience that imitates the way the brain learns from experience. The Deep Feedforward Network, is a composition of functions called layers. They are interconnected in the sense that the information flows always forward, from input to output through a number of hidden layers without feedback or cycles among them. The DFFN is a universal approximation function $Y' = F(X, \theta)$ which learns the parameters to minimize the general optimization problem defined in 1. The advantages of these kind of models is their theoretical flexibility fitting any kind of data only limited by the computational resources. This property is described by the Universal Approximation Theorem [4]. A DFFN can be expressed as a composition of functions as follows

$$\mathbf{F}(X; \Theta) = f^r(, \ldots, f^2(f^1(X; \theta_1); \theta_2), \ldots, ; \theta_r), \tag{6}$$

where X is the input vector, and $\Theta = \{\theta_i\}_1^r$, is the set of parameters of each layer to learn. This function is composed by tree kind of layers: the input, the hidden and the output layer.

The input layer receives a vector X of independent variables. The hidden layers are represented from the first f^1 to the penultimate f^{r-1} function. Each function process the input, and the output layer f^r returns the final processed vector Y'.

Each layer has a number of functions inside called activation functions, which are the simplest unit processing of the network, and each one is denoted as $\phi(X, \theta)$. Each activation function maps the input vectors of size n to a scalar output. Each layer produces an intermediate output vector with the same size according to the number of neurons in the current layer. For the suitability function application, it is possible to fit the data directly without an explicit separation in groups or levels as it does MLLR (using one hot encoding), neither to have an explicit gravitational expression (Equations 4 and 3). The disadvantage of this approach is the loose of explainability about the relations among data variables and the potential overfitting.

5 Experiment and Results

We use the land use change of development detected from year 2000 to 2010. We exclude water bodies, protected natural areas and the already urbanized land (White regions observed in Figure 1b). The set is composed by a total of 7,301,093 vectors excluding the development at year 2000, water shapes, natural areas and industrial parks where the urban development is not possible. These vectors are classified in two classes according to the urban development change. The class 0 containing the 93% of vectors associated with undeveloped cells and the class 1 containing the remaining vectors associated with land use change. In order to select the training set we proceed as follows:

1. We select a random sample with 70% from the set of all vectors class 1 without repetition.

2. To have a balanced sample of classified vectors, we select a random sample of vectors class 0 with the same number of elements contained in Class 1.

3. The remaining data is used as test set, having a test set size of 6,585,741 vectors for DFFN. For MLLR implemented in Futures, the test was performed with the total of the data since the package does not provide functions to identify the data sample used for training.

For measuring the performance of MLLR, we use the implementation of the multi-level logistic function from FUTURES package in GrassGis [10] to measure the allocation accuracy. We use the maps listed in Table 1. The training data are organized in two groups of cells (or levels) associated with the Leon and Silao Municipalities given by sub-region map in Table 1. Meentmeyer et. al. [9], recommend to set α and γ to 0.5, optimizing the linear model with Laplace approximation implemented in lme4 package [1] found in R. The independent variables are contained in vectors $X_i = \{x_{i,1}, \ldots, x_{i,h}\}$ where the number of independent variables is $h = 11$ since we use the variables associated with the first 11 maps presented in Table 1. The municipality is a categorical variable taken from map 12 in Table 1 used as the multi level variable j with the same number of values than the number of regions (e.g., $j \in \{0, 1\}$ because we have two regions defined by Leon and Silao Municipalities).

In order to measure the DFFN performance, we train two model variations testing the dependency of pressure variable γ estimated with Equation 3: DFFN(A) and DFFN(B). DFFN(A) does not require the α parameter but it still requires to experiment manually with γ. For training DFFN(A), the input vector with 11 independent real variables is extended with a codified categorical variable to be fitted directly by the DFFN. Therefore the independent vector of variables is defined as

$$X = \{\{x_1, x_2, \ldots, x_h\}, \{M_j\}\}, \tag{7}$$

where M_j has the codification for the categorical region information associated with Leon, and Silao regions (map 12) with the code vectors $M_1 = (0, 1)$ and $M_2 = (1, 0)$ respectively. This method is called one hot encoding and the code is attached to vector of independent variables generating an input vector of size $h + 2 = 13$.

For DFFN(B) we remove the pressure map to eliminate the trial - error step for γ parameter selection, and build a DFFN driven completely by the optimization process to estimate the best DFFN parameters.

We configure DFFN(A) and DFFN(B) with 3 hidden layers. Each one has 15 rectified unit (relu) activation functions. For DFFN(A), the input layer receives an input vector of size $h + 2$, and for DFFN(B), the input layer receives a vector with 1 less element by eliminating the pressure map. For both DFFN, the output is a probability produced by a logistic function in the range $[0, 1)$ that represents a land use suitability with a probability associated to input vector X_i.

To evaluate the test set, we use the probabilities assigned to each input vector of being classified as urban or non urban given by a logistic function. From the estimated output of test set, we select the 357,676 input vectors (the total of developed cells found in test set) with the highest probability of being class 1, the remaining vectors are set to 0. We perform the classification in this way because in practice the estimation of urbanized cells is given by a regression model.

Table 2: Allocation Accuracy of each model. TP: True Positives, TN: True Negatives, FP: False Positives, FN: False Negatives, TPR:True Positive Rate, TNR: True Negative Rate, MCC: Mathews Correlation Coefficient.

	DFFN(A)	DFFN(B)	MLLR		DFFN(A)	DFFN(B)	MLLR
TP	123692	120203	366021	TPR	**0.8069**	0.7841	0.7097
TN	6402851	6399362	6635747	TNR	**0.9953**	0.9948	0.9779
FP/FN	29599	33088	149671	MCC	**0.8023**	0.779	0.6877

The training was performed with MLPClassifier from the sklearn neural network Python 2.7 package. We train the DFFN with the Adam algorithm, that is a stochastic gradient descent optimization function able to scale with a reasonable high number of vectors.

Next we report in Table 2, the True Positives, False Positives, False Negatives, True Negatives (left column), required to compute the Mathews Correlation Coefficient (MCC) to evaluate the models, and also the complementary True Positive Rate and True Negative Rate coefficients found in right column. MCC is accepted as a balanced coefficient used in binary classes with very different sizes like this case. MCC returns a value between $[-1, 1]$. When $|\text{MCC}| = 1$ means perfect classification agreement or disagreement depending on the sign. If $\text{MCC} = 0$, the proposed classification model is equivalent to use a uniform random number as a prediction model.

Sorting by higher MCC, the Table 2 shows in bold numbers that DFFN(A) has the best Correlation coefficient with respect to DFFN(B) by improving 3%. The second best is DFFN(B) with respect to MLLR by improving the MCC by 13.2%. The same order applies for TPR and TNR.

6 Conclusions

We conclude that DFFN is a good candidate to replace MLLR when the allocation accuracy prediction is required over explicative models like MLLR. The input vectors of independent variables can include categorical and numerical information to train the neural network using codification methods. With this comparison we show that is possible to have a framework based on Machine Learning that helps to avoid manual selection of parameters, and also to have a model that fits well nonlinear relations among the different kind variables.

The disadvantage of this approach, is the difficult interpretation, since the gravitational Equations 3 and 5, are replaced by complex weighted neural connections. Something similar happens with the explainability using linear correlations. Linear correlations considered by MLLR, are replaced also by the neural structure that automatically learns the association of the independent parameters with the output. As future work we propose so far to extend the experiments with the same conditions for both approaches to measure the real improvement of DFFN compared with MLLR, and to consider the inclusion of the urban density, used by the MLLR to produce the pressure map.

Acknowledgments: This project has been supported by Centro de Investigacion en Ciencias de Informacion Geoespacial, CentroGeo.

References

[1] Douglas Bates, Martin Maechler, Ben Bolker, and Steven Walker. *lme4: Linear mixed-effects models using Eigen and S4*, 2014. R package version 1.1-7.

[2] Christa Brelsford, Taylor Martin, Joe Hand, and Luís MA Bettencourt. Toward cities without slums: Topology and the spatial evolution of neighborhoods. *Science advances*, 4(8):eaar4644, 2018.

[3] K. C. Clarke, S. Hoppen, and L. Gaydos. A self-modifying cellular automaton model of historical urbanization in the San Francisco Bay area. *Environment and Planning B: Planning and Design*, 1997.

[4] Balázs Csanád Csáji. Approximation with artificial neural networks. Master's thesis, Faculty of Sciences Eötvöts Loránd University Hungary, Hungary, 2001.

[5] Aaron Dushku and Sandra Brown. Spatial Modeling of Baselines for LULUCF Carbon Projects: The GEOMOD modeling approach. *October*, 2003.

[6] J Ronald Eastman. *IDRISI Selva Manual - Guide to GIS and Image Processing.* IDRISI Production, 2012.

[7] GRASS Development Team. *Geographic Resources Analysis Support System (GRASS GIS) Software, Version 7.2.* Open Source Geospatial Foundation, 2017.

[8] INEGI. Datos. http://bit.ly/2VviieO. Last Visited March 2019.

[9] Ross K. Meentemeyer, Wenwu Tang, Monica A. Dorning, John B. Vogler, Nik J. Cunniffe, and Douglas A. Shoemaker. FUTURES: Multilevel Simulations of Emerging Urban-Rural Landscape Structure Using a Stochastic Patch-Growing Algorithm. *Annals of the Association of American Geographers*, 2013.

[10] A. Petrasova, V. Petras, D. Van Berkel, B. A. Harmon, H. Mitasova, and R. K. Meentemeyer. Open source approach to urban growth simulation. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLI-B7:953–959, 2016.

[11] Brian Pickard, Joshua Gray, and Ross Meentemeyer. Comparing quantity, allocation and configuration accuracy of multiple land change models. *Land*, 6(3), 2017.

[12] Secretaría de Desarrollo Agrario Territorial y Urbano. Programa nacional de desarrollo urbano 2014-2018, 2014. [Access Sept. 2019].