EPiC Series in Engineering

Volume 3, 2018, Pages 1557–1567

HIC 2018. 13th International Conference on Hydroinformatics

A cellular automata urban growth model for water resources strategic planning

Dionysios Nikolopoulos[1], Konstantina Risva[1], Christos Makropoulos[1]

[1] National Technical University of Athens, Athens, Greece
nikolopoulos.dio@gmail.com, conrisva@gmail.com, cmakro@chi.ntua.gr

Abstract. The alarming rate of urbanization poses immediate problems to water resources management, mainly, but not limited to water supply, flood risk management, wastewater treatment and water quality control. Ideally, strategic planning of water systems should be fully aware of the prospects of future urban growth in order to maintain high reliability of services provided and satisfy customers in the long term. Typically, urban growth is handled in a static manner via the development of future scenarios based on previous urban planning studies. Generally, these scenarios focus solely on population increase and ignore the spatial allocation dynamics. Modern urban water strategic thinking needs to incorporate robust tools and methodologies in management practices, able to predict and quantify the outcome possibility of future urban growth. To cope with the aforementioned challenge, this study proposes a novel cellular automata urban growth model as well as, a supplementary remote sensing methodology to preprocess input data.

Keywords: Urban growth, Cellular Automata, Monte Carlo calibration, Urban surfaces identification

1 Introduction

Urban growth is an inherently complex phenomenon. It can be considered as a system of physical expansion and functional transitions. It changes over time through interaction with many inter-related, mostly unknown, drivers and stimulant factors. These interactions create an open, non-linear, dynamic and emergent system. Thus, it is fundamentally difficult, if not entirely impossible, to accurately describe (i.e., model) and predict urban growth. The focus of modelling urban growth should be on creating plausible results and models with good explanatory power, able to simulate and explore the complex urban dynamics, rather than precisely pinpoint future urban locations.

A dominant family of such models is based on Cellular Automata (CA). CA's divide the area of interest in discrete, self-organized cells and operate by applying simple rules defining local interactions among neighbouring cells in each simulation step. The vast number of local interactions can result in a complex and dynamic global behaviour, especially when stochastic rules and disturbances are introduced to the model. Their simple in defining, yet complex in terms of its outcomes, nature makes CA-based models suitable for urban growth simulation and prediction. The literature [1, 2] contains many model sub-types and hybrid interdisciplinary approaches that have been tested extensively in various and diverse cases.

In this work, we have developed a robust, flexible, yet parsimonious (in data requirements) CA model as part of a general methodological framework aimed to assist urban water strategic planning. Briefly, the proposed two-state (urban and non-urban land uses) CA model is conditioned (i.e., constrained) on external drivers and accounts for the allocation dynamics variability via stochastic internal mechanisms. Furthermore, special attention has been given in the modularity of the model in order to provide the means for its straightforward extension and reproducibility. Finally, emphasis has been given on optimizing the simulation speed (i.e., computational effort), a common weakness of most CA models.

2 Introducing DESTICA: An Exogenously Constrained STochastic Cellular Automata model with Descriptive Internal mechanisms

DESTICA is a model comprised of two main modules, an external subsystem that generates the number of new urban cells in the next simulation step, making the model constrained to exogenous drivers, and an internal subsystem that allocates these cells spatially using stochastic mechanisms with descriptive properties. Each subsystem is calibrated to real data before issuing an urban growth prediction. The simulation timestep of DESTICA is annual, as typically new housing estate needs a substantial period of time to develop.

2.1 External subsystem: constraints to exogenous drivers

The external subsystem requires two timeseries as inputs, which are handled as exogenous drivers for urban growth. For the subsystem's calibration, historical timeseries are required in order to replicate as output historical urban growth patterns (i.e. annual timeseries of new urban cells). After calibration, the subsystem can generate future number of urban cells with scenario-based driver predictions. We propose the use of two typical drivers of urban growth as indicated in literature (e.g. [5]): annual GDP and population timeseries. These model inputs should be highly correlated to historical urban growth after some pre-processing. Absolute driver values or relative year over year changes can be unsuitable for calibration (e.g. vary significantly each year) or weakly correlated to urban growth dynamics. In this manner, pre-processing can include transforming absolute values to annual rates of

change or rate of change over a greater period e.g. 5-year rate of change to account for "lag" of the urban system to respond to a driver's change. There is flexibility in assigning different types of drivers as inputs in other case studies, as the external subsystem employs four parameters (a, b, c, d) that transform non-linearly the annual input information to new cells.

The first two parameters (a, b) form a linear model that relates relating new cells Δx with the inputs I_1 and I_2 in step t. Moreover, the urban growth is correlated with the previous step, to account for parsimony and "inertia" in changing cells via c. Parameter d transforms the model with a power law. In addition, a stochastic component can be added to the equation, either after residual analysis or arbitrarily constructed to explore uncertainty and urban system dynamics. The external subsystem in each step t can be summarized by the following equation:

$$\Delta x_t = (aI_1 + bI_2 + c\Delta x_{(t-1)})^d \qquad (1)$$

The parameters are calibrated against historical number of cells changed in each step with the Nash - Sutcliffe Efficiency metric as it gives a qualitative estimation of performance against using the trivial model of generating each step the mean number of observed urban cells changing through the calibration period of n years., i.e. using a steady rate of urban change. In eq. 2 O denotes the timeseries of observed urban changes.

$$NSE = \frac{\sum_{t=1}^{n}(\Delta x_t - \mathrm{O}_t)^2}{\sum_{t=1}^{n}\left[\Delta x_t - \frac{\sum_{t=1}^{n}\mathrm{O}_t}{n}\right]^2} \qquad (2)$$

2.2 Internal subsystem: spatial allocation of cells

The internal subsystem is responsible for the spatial allocation of the new Δx generated cells from the external subsystem in each time step t. The internal subsystem is inherently dynamic as the allocation of cells is stochastic and based on a suitability factor estimation technique (for example [6]), affected by an ensample of mechanisms. Despite its nature, the spatial allocation module of DESTICA is data-parsimonious. Absolute data requirements are a "starting point" urban state in binary form, formed by dividing the region into urban cells and non-urban cells, the topographical slope of the cells derived from a suitable DEM (even from SRTM if not available otherwise), a collection of points of interest that according to the

modeller attract urbanization (e.g. centroids of towns, villages, airports, university campus etc.) and the road network which can be digitized if not available/open from the respective municipality archives.

Three mechanisms affect organic urban expansion (outwards) while the forth induces spontaneous growth (new clusters). Thus, in each step t, a portion of the Δx_t cells denoted ΔE_t corresponds to organic expansion and a portion ΔN_t forms new urban clusters. The ratio between them, as shown in eq. 3 and 4, is calibrated by two parameters signified SC and SG, which respectively describe the percentage of new cells that become the centroid of a new cluster and the percentage of cells that form the immediate neighbours to these clusters, controlling cluster size. For example, low SC and SG values means that most of the urban growth occurs at the "urban edges", while high SC and low SG denotes a scattered urban growth.

$$\Delta N_t = \Delta x_t \times (SC + SG) \quad (3)$$

$$\Delta E_t = \Delta x_t - \Delta N_t \quad (4)$$

The suitability factor for the allocation of ΔE_t cells depends on the weighted factors (eq. 9) derived from the following mechanisms:

The first is an edge expansion mechanism, where cells adjacent to or near urban cells, as defined by a Moore neighbourhood, are more suitable for transition. Suitability of a non-urban cell in position (i, j) in the current step t, symbolized depends on an inverse distance power law, where ρ is the minimum distance to an urban cell counted in intermediate number of cells and d a non-negative calibration parameter defining the rate at which the adjacency effect diminishes. Thus, this suitability ranges from 0 to 1.

$$SFm_{(i,j,t)} = \frac{1}{\rho^d} \quad (5)$$

The second mechanism is slope-dependent. A fuzzy inference system (FIS) is formulated that translates slope input of cell $C_{(i,j)}$ to slope-suitability $SFs_{(i,j)}$ by fuzzy-mapping slope value groups to suitability ranging from "very bad" to "very good", translated to [0,1] suitability score.

$$SFs_{(i,j)} = FIS(C_{(i,j)}) \quad (6)$$

The third mechanism relates road accessibility in the form of cost distance surfaces CDS_x, to every point of interest p_x, with the attractiveness of each point in terms of

population size in year t, marked $POP_{x,t}$ (or another case-dependent weighting factor). Cost is calculated from the least time measured in minutes needed to transverse from a cell to a particular point of interest x, by assigning to each cell $C_{(i,j)}$ the normal transport speed of the major type of road network overlapping it (e.g. 80 km/h in a highway), while cells not overlapping with elements of the road network are assigned the normal walking speed of 5 km/h. A modified gravity model is used to calculate the accessibility - suitability factor SFr_t as shown in eq. 7 and 8. Calibration parameter e of weighting factor $W_{x,t}$ transforms non-linearly the relative importance of population among the different sized points of interest (i.e. how much more "important" is a city to a smaller town in the same region). The non-negative calibration parameter λ implements a rule of distance decay. If needed, a bias can be added to any point of interest to further increase its importance.

$$W_{x,t} = \frac{\ln(POP_{x,t})}{\ln(e)} \qquad (7)$$

$$SFr_t = \sum_{x=1}^{p} \sum_{t=1}^{n} \frac{W_{x,t}}{CDS_{x,t}^{\lambda}} \qquad (8)$$

These three mechanisms are normalized in every step t to 0-1 (denoted by prefix N) in order to form the total suitability factor of each cell by utilizing the equation 9. Parameter w_y denotes the respective weight of the mechanism in the calculation of the suitability factor.

$$SF_t = w_1 NSFm_t + w_2 NSFs_t + w_3 NSFr_t \qquad (9)$$

Every yearly step t, after calculating the suitability factor SF_t of every cell, each cell is tested against a randomly generated number from a "choice function" (which is essentially a beta probability distribution with two calibration parameters a and β) reflecting uncertainty and public disposition against the estimated suitability factor. Some choices for urbanizing a cell are not truly rational e.g. for the reason of already owning a parcel of land. A stochastic ranking procedure is implemented via the subtraction of the randomly generated numbers from the suitability matrix. The ΔE_t cells with the highest score transit to urban state, as represented in eq. 10 and 11.

$$SFc_t = SF_t - PDF_{beta}(rand) \qquad (10)$$

$$C_{transition}(t) \in C \mid SFc_{(i,j),t} \in \{SFc_{rank}(1,t), SFc_{rank}(2,t), \ldots, SFc_{rank}(\Delta E_t) \quad (11)$$

Some cells, ΔS_t (number defined from the SC percentage), that are not near urban cells (distance defined by a cut-off threshold, e.g. more than five cells) change state to urban, based solely on the suitability due to slope and road accessibility. Subsequently, these cells form new urban cluster centroids and some cells, ΔG_t in their immediate neighbourhood change state as defined by SG percentage. This immediate neighbourhood of a cluster centroid is symbolized by $\hat{G}_{(i,j)} = 1$. These changes are shown in the following equations. The same stochastic ranking technique is used to select the ΔG_t fittest cells for urbanization due to the spontaneous urban growth mechanism.

$$SF'_{(i,j),t} = SF_{(i,j),t} + w_3 \forall C_{(i,j)} : NSFm_{(i,j)} = 0 \quad (12)$$

$$SFc'_t = SF'_t - PDF_{beta}(rand) \quad (13)$$

$$C'_{transition}(t) \in C \mid SFc'_{(i,j),t} \in \{SFc'_{rank}(1,t), SFc'_{rank}(2,t), \ldots, SFc'_{rank}(\Delta S_t) \quad (14)$$

$$SFG_{(i,j),t} \begin{cases} 1 - PDF_{beta}(rand) & \text{when } \hat{G}_{(i,j)} = 1 \\ 0 & \text{when } \hat{G}_{(i,j)} = 0 \end{cases} \quad (15)$$

$$C''_{transition}(t) \in C \mid SFG_{i,j,t} \in \{SFG_{rank}(1,t), SFG_{rank}(2,t), \ldots, SFG_{rank}(\Delta G_t) \quad (16)$$

2.3 Calibration of internal subsystem

The parameters of the internal subsystem (ten in total) are calibrated against the similarity of the output to the real urban changes across a specific timespan, using three weighted metrics. The metrics used are the modified Cohen's Kappa coefficient of agreement K_{sim} [3], a shape index and the agreement by number of clusters. Specifically, K_{sim}, which is based on the contingency matrix of the classes, is used to determine solely the performance of the model in the spatial allocation because DESTICA is calibrated using the observed instead of the simulated number of cells changed [7]. In eq. 17 p_o represents the observed fraction of agreement between the allocation of the classes [urban, non-urban] and $p_{e(Transition)}$ represents the expected fraction of agreement, given the sizes of the class transitions to other classes. Allocation agreement alone is not a representative indicator of performance [7]. Realistically, pixel per pixel accuracy in allocating urban change cannot be very

high, considering the usual size of a suburban area (large), the size of urban changes over a short time window (small) and the related degrees of freedom in cell allocation. Also, the model dynamics allow a variety of different emergent behaviors derived from dissimilar parameter sets to achieve similar performance. Therefore, the shape and cluster number metrics support performance evaluation by giving an estimation of "compactness", dispersion and patterns of urban growth. Shape is calculated as the ratio of the total area A of the k clusters to the total perimeter P. Both shape and number of clusters metrics are compared to the observed indices of the urban area, $ShapeRef$ and $ClustersRef$. The comparison is made between the output of DESTICA (last simulation step in the calibration process) and the reference urban state. The optimization problem is defined as the minimization of the objective function $objF$ in eq. 19, using different weights β_i.

$$K_{sim} = \frac{p_o - p_{e(Transition)}}{1 - p_{e(Transition)}} \tag{17}$$

$$Shape = \frac{\sum_{k=1}^{N} A_k}{\sum_{k=1}^{N} P_k} \tag{18}$$

$$objF = \beta_1 \left| \frac{ShapeRef - Shape}{ShapeRef} \right| + \beta_2 \left| 1 - K_{sim} \right| + \beta_3 \left| \frac{ClustersRef - Clusters}{ClustersRef} \right| \tag{19}$$

Due to the stochastic nature of the model, a Monte-Carlo technique is applied in combination with a genetic algorithm for optimization. For each individual of the population, the model is executed a pre-specified number of times for the same parameters (e.g. 100, determined by the available processing capability). The performance is evaluated each time and the median value of the $objF$ is utilized to assess the score of the individual.

3 Identification of historical urban growth

Alongside DESTICA, an urban classifier model is developed in order to enable it to use open data from remote sensing sources, such us Landsat images. Landsat is the longest running open data mission, and therefore is a suitable source for generating a large enough timeseries to determine, via remote sensing historical, urban change in an area. In addition, the medium resolution of the produced data (ca. 30 m × 30 m) is sufficient for urban identification. Other data collections with lower resolution (e.g. MODIS, ca. 250 m × 250 m) provide a coarser approach to the classification. High resolution data (e.g. Worldview ca. 0.5 m × 0.5 m) are typically commercial and

recent. Therefore, their timeseries is limited, while processing time rises exponentially with resolution.

Coupling remote sensing with an urban growth model is a key point of our methodology, as suitable and accessible data are in many cases scarce and should be inferred by other means. However, identification of urban areas from satellite data is an inherently difficult task. We developed an urban classifier model that uses an artificial neural network (ANN) structure utilizing not only multispectral bands as inputs, but also the texture info. We constructed various sets of predictors for ANN input that included: a) only spectral bands except panchromatic, thermal and special bands of Landsat 8 (i.e. for R, G, B, IR, MIR1 and MIR2 bands - 6 predictors), b) spectral bands plus 20 available Haralick-GLCM texture indices for every spectral band for moving windows of 3×3, 5×5, 7×7 pixels (126 predictors for each moving window), c) spectral bands plus selected contrast, dissimilarity, energy and entropy texture info for the bands R, IR, MIR1, MIR2, for moving windows of 3×3, 5×5, 7×7 (22 predictors for each), d) spectral bands and multivariate variogram methods for texture info (7 predictors), where multivariate info between bands is calculated either by Euclidean, Mahalanobis or spectral angle distance [4], for 3×3 moving window. We also use a Monte-Carlo scheme to account for uncertainty in selecting samples. Control samples for calibration, validation and testing are randomly selected from a predefined pool (from basemaps, Google Earth etc.) and the classification performance is evaluated 1000 times. The set of predictors with the best distribution of performance is used to select a classifier type that classifies the most recent satellite image, because the control samples are readily available only for recent images in most cases. The output of the classifier is then combined with an object-oriented hierarchical (in a temporal manner) classification methodology to derive the required historical land-use change timeseries from the other preceding Landsat images. The latter is required to calibrate the parameters of DESTICA model. In this manner, we use an iterative process of employing the multiresolution segmentation algorithm of eCognition software to define objects in each year, and test resemblance to the previous year to identify any changes.

4 Case study & results

The above methodology is applied to the greater area around Rethymno city in Crete, Greece. The area includes various settlements (and a university campus) that differs significantly in size and patch density. Centroids are used as points of interest. The mixing of different land uses greatly increases the difficulty for both the urban classifier model and DESTICA. Also, historical data are scarce, incomplete and inaccurate. This paradoxically makes it an ideal case to demonstrate the robustness of the new methodology.

In order to calibrate the ANN urban classifier, 1390 control samples are selected from the area mainly using Google Earth, comprised of 729 non-urban and 661

urban samples. Classification is performed on a Landsat 8 image of July 2016. Monte Carlo analysis is performed utilizing 1000 random subsets of the samples split in 70% calibration, 15% validation and 15% testing. The predictor set that performs better is the one with the 7 predictors comprised of the 6 spectral bands and the multivariate texture using spectral angle distance. As shown in Figure 1, the mean classification error is 5.24%, while standard deviation is 0.94%. The best classifier, which is subsequently used to classify the whole Landsat image, has a 2.73% classification error.

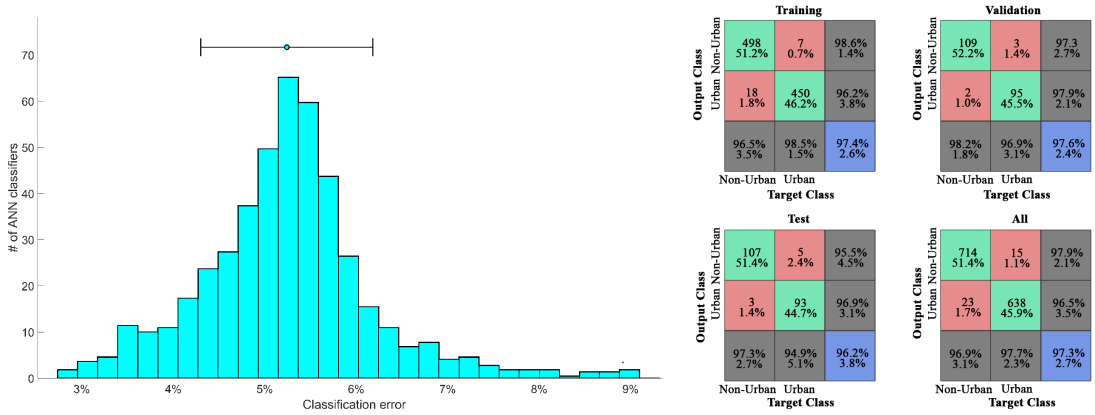

Figure 1 Left: Classification error distribution of the best predictor set (spectral bands and spectral angle distance multivariate texture). Right: Confusion matrices of the best classifier.

Classified images are presented in Figure 2. The urban classifier model provides the initial state (top left) of year 1984 and the end state of year 2016 (top right) based on the respective Landsat 5 and 8 images. In order to demonstrate DESTICA flexibility and magnify the attractiveness of two specific points of interest, 'University Campus' and 'Old Town', two more bias weights, p_1, p_2, are added to eq. 7. DESTICA simulation from years 1984 to 2016 after the calibration of the 12 internal model parameters $(SC, SG, d, e, \lambda, w_1, w_2, w_3, \alpha, \beta, p_1, p_2)$ can be visualized (bottom right) along with the heatmap of urban state probability, after 100 model trials (bottom left). Table 1 presents the parameter values and the evaluation of metrics $Shape, Ksim, Clusters$. The respective weights for calibration are $[\beta_1, \beta_2, \beta_3] = [0.45, 0.05, 0.5]$.

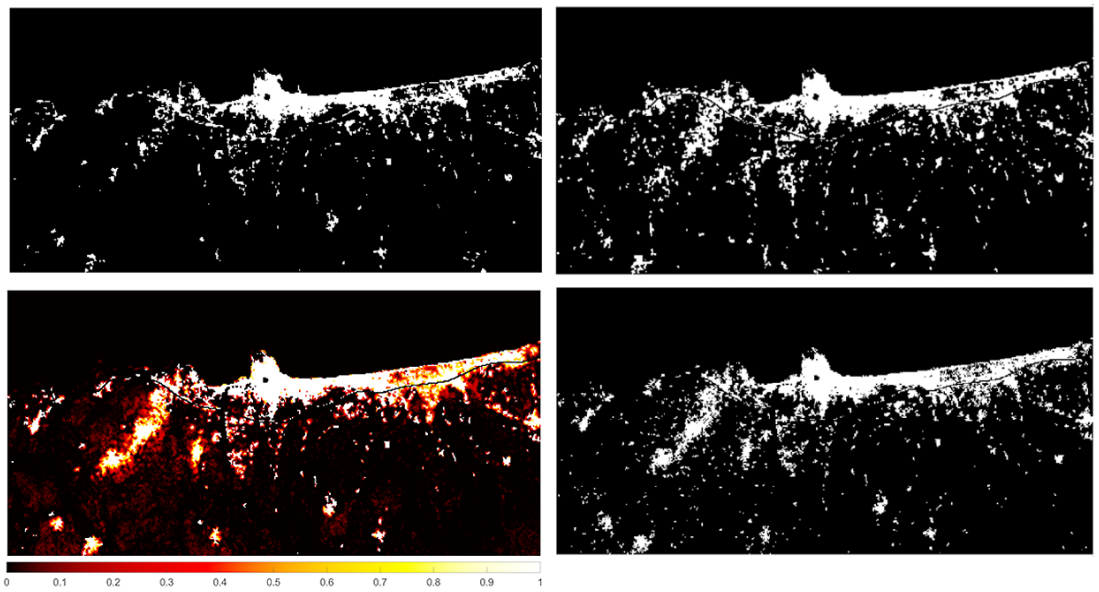

Figure 2 Top left: initial binary state of year 1984 (□: urban cells, ■: non-urban cells). Top right: real end state of year 2016. Bottom right: simulation end state of year 2016. Bottom left: probability heatmap of cells being urban at simulation end state after 100 model trials.

Table 1 DESTICA internal model parameters and metrics evaluation

SC	SG	d	e	λ	w_1	w_2	w_3	α	β	p_1	p_2
0.106	0.113	1.765	3.237	1.068	0.645	0.068	0.287	215	119	4580	3320
	Shape				K_{sim}			*Clusters*			
	1.76 (*ShapeRef* : 1.74)				0.21			532 (*ClustersRef* : 530)			

The external subsystem is calibrated with inputs I_1, I_2 representing the annual growth rate of GDP with lag of three years and the average 5-year growth rate of Rethymno population. The parameter values of $[a,b,c,d]$ are [24.6, 262.4, 0.10, 1.56]. The NSE value of 0.70 signifies good performance of the external model in simulating annual urban changes for the specified period (1984-2016). DESTICA is fully calibrated and could be used for a future urban state prediction for the area of Rethymno. Figure 3 presents an example prediction of urban state at year 2040 and an application in defining flood risk in the eastern part of the city for a storm of $T = 100$ years return period. New urban patches in the areas marked with deep blue colour are the ones at higher risk.

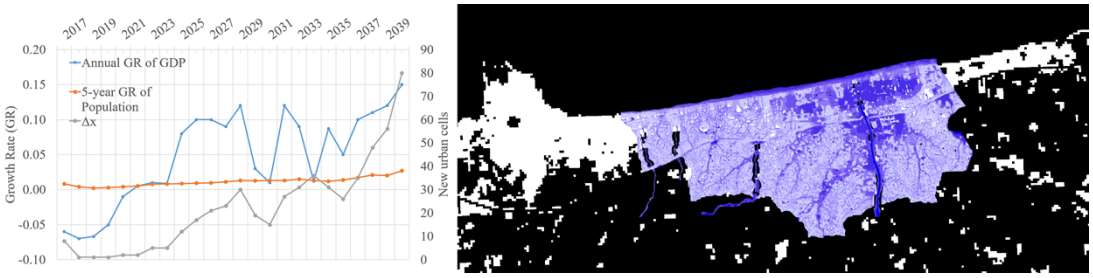

Figure 3 Left: prediction example of new urban cells (grey) until 2040 using "dummy" GDP and population predictors. Right: the new urban state at the eastern part of the city and flood depths in blue tone colormap.

5 Conclusions

The results are very promising, indicating that the methodology is capable of both identifying historical urban growth from open remote sensing data using the urban classifier model and simulating the complex nature of urban growth, even in unfavorable conditions, via the use of DESTICA. The hypothetical future water resources-related example highlights the utility of the proposed methodology in the context of modern urban water strategic planning.

Reference

[1] X. Li, P. Gong, Urban growth models: progress and perspective, Int. J. Earth Sci. 61 (2016) 1637-1650.

[2] I. Santé, A. García, D. Miranda, R. Crecente, Cellular automata models for the simulation of real-world urban processes: A review and analysis, Landscape Urban Plan. 96 (2010), 108-122.

[3] J. Van Vliet, A. Bregt, A. Hagen-Zanker, Revisiting Kappa to account for change in the accuracy assessment of land-use change models, Ecol. Model. 222 (2011), 1367-1375.

[4] J. Zhang, P. Li, J. Wang, Urban Built-Up Area Extraction from Landsat TM/ETM+ Images Using Spectral Information and Multivariate Texture, Remote Sens. 6 (2014), 7339-7359.

[5] K. C. Seto, R. Sánchez-Rodríguez, F. Michail, The New Geography of Contemporary Urbanization and the Environment, Annu. Rev. Environ. Resour. 35 (2010), 167-194.

[6] D. Bouziotas, E. Rozos, C. Makropoulos, Water and the city: exploring links between urban growth and water demand management, J. Hydroinform. 17 (2015), 176-192.

[7] D. Nikolopoulos, Development of geospatial urban growth models supporting urban water strategic planning: the case study of Rethymnon, Crete, Master Thesis, NTUA, Athens, 2017.