Urban Dynamic Expansion Computer Simulation of RF-NH-CA Model Considering Neighborhood Heterogeneity

Junyi Li1,2, Minghao Liu1,2,*, Tianlin Liu1,2, Lei Jing1,2

1College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, China
2Spatial Information Research Center, Chongqing University of Posts and Telecommunications, Chongqing, China

*Corresponding author: liumh@cqupt.edu.cn

Abstract. "Neighborhood" as the principle of "the closer the distance, the more relevant the attributes", is often used as a key driving factor for the urban dynamic modeling of cellular automata; however, the current implementation of the "neighborhood" idea is mostly adopted Mean probability method. This method affects the accuracy of urban dynamic simulation to a certain extent because it ignores the spatial heterogeneity of neighboring cells. Based on the random forest method to evaluate the suitability probability of land use, this study uses the intensity gradient change characteristics of the luminous data to endow the traditional neighborhood cell heterogeneity characteristics, and builds a random forest neighborhood heterogeneity CA model (Random forest Neighborhood Heterogeneity Cellular Automata, RF-NH-CA), and verified the effectiveness of the model by simulating the changes in urban land use in the 21 districts of Chongqing’s main city from 2010 to 2017 through a multi-scheme comparative experiment. The results showed that the overall simulation accuracy of the RF-NH-CA model reached 97.59%, and the Kappa coefficient reached 0.7434; compared with the traditional models RF-CA, ANN-CA and Logistic-CA, FoM increased by 0.0274, 0.0383, 0.0579, respectively. The Kappa coefficient increased by 0.0162, 0.0229, 0.0351 respectively. Studies have shown that giving the neighborhood cell heterogeneity through luminous data has played a role in improving the accuracy of land use simulation, which is more in line with the real urban expansion.

Keywords: Cellular automata; urban expansion; night light data; neighborhood heterogeneity.

1. Introduction

Cellular automata based on the idea of "complexity comes from simplicity, local influences the whole world" are widely used in the spatiotemporal modeling of complex phenomena in the fields of urban expansion simulation, climate change impact assessment, and air pollution diffusion simulation [1-3]. In recent years, it has become an important tool in the simulation of urban dynamic changes due to its advantages of discrete time and space, "bottom-up", and simple modeling [4,5].
At present, domestic research on cellular automata is mostly focused on the extraction of rules and the exploration of neighborhoods [6]. From the perspective of cellular automata rule extraction, in recent years, machine learning and deep learning methods have been widely used in cellular automata rules to become a new trend, such as ANN-CA (Artificial Neural Network-CA) model, Logistic-CA model [7], DT-CA (Decision Tree-CA) decision tree cellular automata model [8], MCR-CA (Minimal Cumulative Resistance-CA) cellular automata model based on minimal cumulative resistance [9], etc. And other urban expansion models. From the perspective of neighborhood research, “neighborhood”, as the specific application of the first law of geography on the principle of “the closer the distance, the more relevant the attributes”, is often used as a key driving factor for the dynamic modeling of the city by cellular automata. In recent years, in addition to discussing the optimal neighborhood scale [10] and irregular neighborhoods [11-12], few scholars have studied the neighborhood heterogeneity of cellular automata. Among them, the intensity of night lights is due to a certain extent, it indicates that the city’s economic development and regional differences [13] are extracted for urban economic modeling; moreover, compared with the DMSP-OLS night light data, it is carried by the newly launched observation satellite Suomi NPP in 2011. The night light remote sensing image data acquired by Visible Infrared Imaging Radiometer Suit (VIIRS) has high spatial resolution and a wider radiation detection range.

In view of this, this paper uses the characteristics of the intensity gradient of the luminous data to introduce the NPP-VIIRS night light data into the cell neighborhood, and endows the traditional neighborhood with the cell heterogeneity characteristics, and constructs the random forest neighborhood heterogeneity CA model (Random forest Neighborhood Heterogeneity Cellular Automata, RF-NH-CA), and through a multi-scheme comparison experiment to simulate the changes in urban land use in the 21 districts of Chongqing’s main city from 2010 to 2017 to verify the effectiveness of the model.

2. Research area and data
Located in southwestern China, Chongqing is one of the core cities in the Chengdu-Chongqing Economic Circle. It is playing a supporting role in the efforts to promote the development of the western region in the new era, playing a leading role in the joint construction of the "Belt and Road", and promoting the Yangtze River Economic Belt. Playing an exemplary role in green development. Specific as shown in Figure 1.

![Figure 1 Schematic diagram of Chongqing metropolitan area](image-url)
The driving force of urban expansion is mainly influenced by three factors: natural environment, social economy and traffic conditions. This study selected driving factors from the above three aspects and constructed a data set containing 19 driving factors. Night light intensity, slope, distance and elevation at the airport, the distance to the subway, the school point density, bank point density, to expressway entrance point density, to train the site density, the density of attraction point, to the main railway distance, distance to water, to the town center distance, the density of related sites to shopping, the distance to the main road, entertainment, hospital, restaurant point density point density Site point density, hotel point density. (Table 1).

### Table 1 Data used in this study

| Name                        | Years          | Resolution | Source                                      |
|-----------------------------|----------------|------------|---------------------------------------------|
| Land use data               | 2010-2017      | 30 m       | FROM-GLC (data.ess.tsinghua.edu.cn)         |
| Slope                       | 100 m          | Calculated by DEM |
| DEM                         | 100 m          | WorldPop (www.worldpop.org)                |
| Night light data (DMSP data source) | 2010 100 m | WorldPop (www.worldpop.org)                |
| Night light data (NPP-VIIRS data source) | 2016 100 m | WorldPop (www.worldpop.org)                |
| Water body                  | 2010           | OpenStreetMap                                       |
| Road network data           | 2010           | Web crawler acquisition                             |
| Point of interest data      | 2010-2017      | Web crawler acquisition                             |

3. **Principles and methods**

3.1. **RF-NH-CA model**

In the urban CA model, the core of the cellular automata model is the extraction of conversion rules. Among them, neighborhood rules play a very critical role in the evolution of the city during the long period of evolution. The conversion rules in this study mainly consider the following four aspects: urban land development suitability $P_0$, neighborhood part $Q_1$, limiting factor $P_2$ and random factor $RA$. $P_0$ in this paper is calculated by using random forest algorithm (RF). Taking into account the level of economic development between urban construction land and the heterogeneity between urban cells caused by regional differences, and in reality, urban expansion is restricted by land use types and the government's decision-making ability. A CA model (Random forest Neighborhood Heterogeneity Cellular automata, RF-NH-CA) model based on night light data (NPP-VIIRS data) was constructed to reflect the heterogeneity of neighborhoods.

3.1.1. **Suitability evaluation**. When using random forest to predict, for unknown categories, the method adopted by random forest is determined by the majority vote of $n$ trees, and finally the classification result is obtained$^{[14]}$, the formula is as follows:

$$H(x) = \arg \max \sum_{j=1}^{n} I(h_j(x) = Y)$$

In the above formula: $\arg \max \sum_{j=1}^{n} I(h_j(x) = Y)$ represents the value of $Y$ when $\sum_{j=1}^{n} I(h_j(x) = Y)$ is the maximum value; $h_j(\cdot)$ represents the $j$ th decision tree Model; $Y$ represents the classification result of the decision tree; $I(\cdot)$ represents the indicator function of the classification result, and $n$ represents the number of subtrees of the random forest.

The probability that a cell $i$ is classified into the k-th category is shown in equation (2):
\[ P_0(i, k) = \frac{\sum_{j=1}^{n} l(h_j(x_i) = y_k)}{n} \]  

(2)

In the urban expansion simulation study, the probability \( P_0(i, k) \) (also representing the urban development adaptability of cell \( i \) at time \( t \)) that cell \( i \) is transformed into a city cell at time \( t \) can be expressed as:

\[ p_0(i, t) = \frac{\sum_{j=1}^{n} l(h_j(x_i) = \text{urban})}{n} \]  

(3)

3.1.2. Restrictive evaluation. Under real circumstances, special land use types will not be converted into cities, such as water bodies and forests will not be converted into cities, so the limiting factors can be expressed as:

\[ P_c(i, t, k) = \text{con}(S_i^k = \text{Landuse}_k) \]  

(4)

In the above formula: \( \text{Landuse}_k \) represents the \( k \)-th land use type. When the cell \( i \) meets the land use type that meets the allowable conversion at time \( t \), the value is 1, otherwise the value is 0;

3.1.3. Random factor. As the city is expanding, there may be some emergencies, such as policy adjustments, the impact of economic development, and the occurrence of random circumstances such as natural factors. Therefore, this model introduces random factors. The specific formula is as follows:

\[ RA = 1 + (-ln y)^a \]  

(5)

In the above formula: \( RA \) is the random factor part, the value of \( y \) is a random number in the range of \((0, 1)\), and the parameter of formula \( a \) controls the degree of influence of the random variable (the value range is an integer from 1 to 10).

3.1.4. Neighborhood factor. The neighborhood effect is an important part of the cellular automata, while the traditional cellular automata only considers the influence of the neighborhood effect in terms of quantity, and does not reflect the fact that the urban construction land cells are affected by the region and economy. Situation and other differences caused by the impact of the difference. As shown in Figure 2, although the layout of the left and right images is different, for the non-urban grid in the center, there are 4 urban grids around. If the traditional neighborhood algorithm is used, the value of the center pixel is only the same as that of the surrounding pixels. The number of grids has nothing to do with the size of the grid value. It ignores the heterogeneity of urban grids caused by geographical, economic and other factors.

In view of the gradual improvement of the spatial resolution of luminous data in recent years, the phenomenon of urban internal light generalization has been suppressed, and the gradient of light intensity at night has gradually increased. The luminous intensity itself reflects the economic development of a region and the possibility of future urbanization to a certain extent. Therefore, with the help of the characteristics of the intensity gradient of the luminous data, it is proposed to introduce the NPP-VIIRS night light data into the cell neighborhood, and use the brightness value of the NPP-VIIRS night light as a measure of the “quality” level and difference of the city grid. The brightness value of the NPP-VIIRS night light data after conversion is used as the spatial heterogeneity parameter.

As shown in Figure 3, for the same neighborhood surrounded by 4 city grids, because the brightness values of the NPP-VIIRS night light data of different city grids are different, it means that the economic development level between the city grids is inconsistent. To a certain extent reflects the spatial heterogeneity of the urban grid, which is more in line with the real situation.

The neighborhood index of traditional cellular automata can be expressed by the following formula:
The neighborhood index calculation formula (7) that introduces night light data to reflect the spatial heterogeneity of neighborhoods is as follows:

\[
Q_i^t = \frac{\sum_{n=1}^{n} l_j \cdot con(S_j=urban) \cdot l_i}{n^2-1}
\]  

(7)

In the above formula, \(Q_i^t\) represents the value of the \(n \times n\) neighborhood of location \(i\) at time \(t\), \(l_i\) represents the NPP-VIIRS night light brightness value after the city grid at location \(i\) is normalized, and \(l_j\) represents location \(j\) NPP-VIIRS night light brightness value (0-1) after normalization of the city grid, \(con(\cdot)\) is a conditional function, \(S_i^t\) represents the state of the cell at position \(i\), if the current cell is a city cell, the value is 1, otherwise the value is 0.

3.1.5. RF-NH-CA model. The conversion probability of cell \(i\) at time \(t\) is:

\[
P_{lk}^t = P_c(i, t) \times Q_i^t \times P_c(i, t, k) \times RA
\]  

(8)

After calculating with the above formula, the probability that the central cell will be converted to the city cell can be obtained, and then the conversion rule can be determined by the principle of selecting the best to control the total amount of city cells that will be converted from the base year to the last year:

\[
TotalNum = \sum_{\tau} stepNum
\]  

(9)
if \( p_k^t \) in \( \bigcup_{\text{stepNum}} \text{DescendingSorted}(p^t) \):

\[
S_k^{t+1} = 1
\]  

(10)

Then update \( Q_k^{t+1} \)

In the above formula, TotalNum is the total number of city cells converted from the base year to the last year; \( T \) represents the number of cycles set during the simulation period; \( \text{stepNum} \) is the number of city cells converted each time, \( \bigcup_{\text{stepNum}} \text{DescendingSorted}(p^t) \) indicate the overall development probability according to your order from largest to smallest, then take the first \( \text{stepNum} \) cell set. At the same time, in order to reflect the differences between the different RF-NH-CA models and the traditional cellular automata CA model, according to the control variable method, the neighborhood part is calculated according to their respective calculation methods, while the other parts remain the same.

3.2. Technical scheme and process

3.2.1. Scheme design. In order to verify whether the introduction of night light data intensity to reflect the heterogeneity of neighboring cells can improve the accuracy of the model, a comparative experiment between the RF-NH-CA model and the RF-CA model is constructed; in order to verify the RF-NH-CA model, the design and construction of the RF-NH-CA model and the traditional urban expansion simulation prevention Logistic-CA, ANN-CA comparative experiment.

3.2.2. Technical process. The model proposed in this paper consists of a random forest model and the introduction of night light data to improve the traditional neighborhood CA model, and its structure (Figure 4). The operation process of the model is as follows: 1) Through the land use data of 2010 and 2017, the urban change data from 2010 to 2017 and the amount of urban expansion are obtained through land use change detection. 2) Randomly stratified sampling of urban change data and driving factor data through Python scripting language, and training and simulation of random forest model constructed through Python to obtain the city's development suitability probability. 3) By normalizing the night light data according to formula (6) and formula (7), the probability value of transforming the non-urban grid of the neighborhood effect into the urban grid domain effect is obtained. 4) Construct a cellular automata model based on night light data through limiting factors, neighborhood factors, urban development suitability, and random factors, and simulate urban expansion and parameter verification through the iteration of cellular automata to obtain simulation results. 5) Model checking. The simulation results of 2010-2017 are compared with the actual land use in 2017, and the accuracy of the model is evaluated by calculating FoM\(^{[15]}\), overall accuracy, and Kappa coefficient, to verify the accuracy of the model, and compare it with RF-CA, ANN-CA and Logistic-CA Compare the models to verify the accuracy of the models.

The specific technical flow chart of the RF-NH-CA model is shown in Figure 5. (1) Firstly, the total conversion number of city cells TotalNum, the number of iterations \( T \), and the number of conversion cells per iteration \( \text{stepNum} \) are calculated through land use change detection; (2) Calculate the neighborhood index and the global development probability by formula (7); (3) Sort the global development probability in descending order, select the first \( \text{stepNum} \) cells for transformation, that is, complete an iteration, update the neighborhood and recalculate the global development probability; (4) Whether to perform the next cycle according to the conditions, if yes, return to step (2), otherwise complete the simulation process to generate simulation results.

The specific technical flow chart of the RF-NH-CA model is shown in Figure 5. (1) Firstly, the total conversion number of city cells TotalNum, the number of iterations \( T \), and the number of conversion cells per iteration \( \text{stepNum} \) are calculated through land use change detection; (2) Calculate the neighborhood index and the global development probability by formula (7); (3) Sort the global development probability in descending order, select the first \( \text{stepNum} \) cells for transformation, that is, complete an iteration, update the neighborhood and recalculate the global development probability; (4) Whether to perform the next cycle according to the conditions, if yes, return to step (2), otherwise complete the simulation process to generate simulation results.
Slope
elevation data
Distance to road
and river data
Points of Interest
GIS database
Random stratified
sampling
Random forest
model (Training)
Random forest
model (Simulation)
Land use data
(Last year)
Land use data
(Base year)
Urban land use
changes data
Land use change
detection
Data processing
Urban Development
suitability
Neighborhood
index
Random factor
Limiting factor
Overall probability
of conversion
Comprehensive
probability calculation
Model validation
And application
RF-NH-CA Model
Total urban sprawl
RF-NH-CA Model

**Figure 4** RF-NH-CA model structure

**Figure 5** Flow chart of RF-NH-CA model
4. Model application and its results

4.1. Model training and parameter setting
First, in order to obtain the global suitability probability, the random forest model needs to be trained. Therefore, through the land use change detection in 2010 and 2017, the urban land use change data is obtained, and the driving factor data is obtained through ArcGIS software processing, and the training sample data is obtained by combining with the land use change data, and then the total sample is drawn through the stratified sampling method 10% of the data is input into the random forest model for tuning and training. There are three important parameters that affect the random forest model: the number of subtrees n_estimators, the maximum depth of the tree max_depth, and the number of driving factors considered when the subtree nodes are split, max_feature.

For the tuning of the n_estimators parameter, the optimal n_estimators is 243; for the search for the max_depth optimal parameter, the final result is when max_depth is 37, the effect is the best. For the search for the optimal parameter of max_feature, the parameter of the optimal max_feature is finally obtained as 4.

At the same time, input all the sample data into the trained random forest model to get the global suitability probability. The random factor part takes into account the small external influence factors, so the random factor parameter a takes a smaller value of 2. The neighborhood effect part is introduced into the neighborhood part by using the normalized NPP-VIIRS night light data of 2016. $Q_i$ is calculated by formula (7), and then the global cell is calculated by formula (8) to develop into a city cell. Probability. From 2010 to 2017, 70,358 new urban cells (TotalNum) were added from 2010 to 2017 through the detection of changes in urban cell land use, and the number of cellular automata cycles (T) is 7, then the number of cells converted in each cycle (stepNum) is about 10051, and the global cell state is updated after each loop.

4.2. Urban expansion simulation results
From the simulation results of various models (Figure 6), Logistic-CA, ANN-CA, and RF-CA are compared with the actual urban distribution of Chongqing 21 districts in 2017, and the RF-CA model is better than Logistic-CA and ANN-CA. A simulation that is more in line with the real situation. Based on the RF-CA model, this paper proposes the RF-NH-CA model that reflects the heterogeneity of the neighborhood of cellular automata (Figure e). Compared with the real city distribution (d), the RF-NH-CA model is better than the RF-NH-CA model. It shows that reflecting the heterogeneity of neighborhoods through night lighting data is more in line with the real urban expansion simulation.
4.3. Accuracy evaluation

It can be seen from Table 2 that compared with the traditional RF-CA model, the overall accuracy of the RF-NH-CA model is increased by 0.0016, the Kappa coefficient is increased by 0.0162, and the FoM coefficient is increased by 0.0274. It shows that the RF-NH-CA model based on NPP-VIIRS night light data to improve neighborhood heterogeneity can simulate a more realistic urban form.

In order to reflect the difference between the RF-NH-CA model proposed in this paper and the traditional urban simulation model, the RF-NH-CA model is compared with the traditional cellular automata model. The experimental results show that: Compared with ANN-CA, the overall accuracy, Kappa coefficient and FoM coefficient of the RF-NH-CA model are improved by 0.0025, 0.0029 and 0.0383 respectively. Compared with Logistic-CA, it increased by 0.0033, 0.0351 and 0.0579, respectively. It shows that the RF-NH-CA model proposed in this paper has better performance for the traditional CA model and for city simulation.

| parameter | RF-LIGHT-CA | RF-CA |
|-----------|-------------|-------|
| OA        | 0.9759      | 0.9743|
| Kappa     | 0.7434      | 0.7272|
| FoM       | 0.3392      | 0.3118|

| parameter | RF-NH-CA | ANN-CA | Logistic-CA |
|-----------|----------|--------|-------------|
| OA        | 0.9759   | 0.9737 | 0.9726      |
| Kappa     | 0.7434   | 0.7205 | 0.7083      |
| FoM       | 0.3392   | 0.3009 | 0.2813      |

5. Conclusion

Aiming at the traditional CA model that only defines the calculation method of the neighborhood quantitatively, it does not reflect the problem of the spatial heterogeneity of urban cells. This paper uses
the gradient change characteristics of the luminous data to introduce the NPP-VIIRS night light data into the neighborhood. Given the heterogeneous characteristics of traditional neighborhood cells, a random forest neighborhood heterogeneity CA model was constructed. From a realistic point of view, Chongqing 21 District was used as an example to simulate the urban expansion of Chongqing 21 District from 2010 to 2017. In the end, the simulation accuracy of the model is verified through the calculation of overall accuracy, Kappa coefficient and FoM coefficient.

The research shows that: 1) Compared with the RF-CA model, the overall simulation accuracy, Kappa coefficient, and FoM coefficient of the RF-NH-CA model that reflects the urban spatial heterogeneity by introducing NPP-VIIRS night light data into the neighborhood RF-CA model, so the CA model considering the heterogeneity of neighborhood is more in line with the real urban expansion simulation. 2) Comparing the constructed RF-NH-CA model with the traditional model, RF-NH-CA has a certain improvement in simulation accuracy, indicating that the constructed RF-NH-CA model is more efficient than the traditional CA model.

Urban expansion change is an extremely complex and dynamic change system, and the driving force of urban expansion areas is also affected by various aspects. Our simulation cannot take all situations into consideration, although this article contributes to driving factors. A certain analysis has been done, but the influence of driving factors in different areas is still not considered, so zoning simulation is the direction that can be explored next. This article uses the traditional 3*3 Moore neighborhood, and does not consider the introduction of night light data into the neighborhood under different types of neighborhoods and neighborhoods of different sizes to reflect the impact of heterogeneity on them; night light data can only reflect the heterogeneity of neighboring cities to a certain extent. In the future, we can explore more suitable factors to reflect the heterogeneity of neighboring cities, and introduce them into the neighborhood to reflect the spatial heterogeneity of neighborhoods. Thereby improving the accuracy of the model.

References

[1] LU Q, CHANG N-B, JOYCE J, et al. Exploring the potential climate change impact on urban growth in London by a cellular automata-based Markov chain model[J]. Computers, Environment and Urban Systems, 2018, 68.
[2] Wei Zongming, Ai Jiaoyan, Li Xiuhua, et al. Atmospheric diffusion model and GIS visualization based on cellular automata[J]. Environmental Engineering Science, 2016, (5): 2527-2534.
[3] QIAO Y F, XUE Y, WANG X, et al. Investigation of PM emissions in cellular automata model with slow-to-start effect[J]. Physica A: Statistical Mechanics and its Applications, 2021.
[4] Zhao Linfeng, Liu Xiaoping, Liu Penghua, et al. Urban expansion simulation and early warning based on geographical zoning and FLUS model[J]. Journal of Geo-Information Science, 2020, 22(03):517-530.
[5] Hu Zui, Wang Hui. Construction and Application of Cellular Automata of Urban Spatial Expansion Constrained by Multiple Factors[J]. Journal of Geo-Information Science, 2020, 22(03):616-627.
[6] AHMED, MUSTAFA, ISMAïL, et al. Measuring the Effect of Stochastic Perturbation Component in Cellular Automata Urban Growth Model[J], 2014.
[7] Li Xia, Ye Jianan. Unit automata CA based on neural network and real and optimized city simulation[J]. Acta Geographica Sinica, 2002, (02):159-166.
[8] XIA L I, YEH G O J I J O G I S. Data mining of cellular automata's transition rules[J], 2004, 18(8):723-744.
[9] Li Xiuquan. Comparative analysis of urban expansion simulation based on MCR and CA models[D]. East China University of Technology, 2017.
[10] FENG Y, LIU Y, TONG X, et al. Modeling dynamic urban growth using cellular automata and particle swarm optimization rules[J]. Landscape & Urban Planning, 2011, 102(3):188-196.
[11] PINTO N, ANTUNES A P, ROCA J. Applicability and calibration of an irregular cellular automata model for land use change[J]. Computers Environment & Urban Systems, 2017, 65:93-102.

[12] BARREIRA-GONZALEZ P, AGUILERA-BENAVENTE F, GOMEZ-DELGADO M. Implementation and calibration of a new irregular cellular automata-based model for local urban growth simulation: The MUGICA model[J]. Environment and Planning B, 2019, 46(2):243-263.

[13] Wang Junhua, Zhang Tingbin, Yi Guihua, et al. Spatial analysis of Sichuan GDP based on DMSP/OLS night light data[J]. Science of Surveying and Mapping, 2019, 44(08):50-60.

[14] Random Forests[J]. Machine Learning, 2001.

[15] PONTIUS R G, BOERSMA W, CASTELLA J C, et al. Comparing the input, output, and validation maps for several models of land change[J], 2008, 42(1):11-37.