Urban land use simulation based on ProWSyn-MLP-CA

Sinan Chen¹,², Minghao Liu¹,²,*, Tianlin Liu¹,² and Junyi Li¹,²
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. In the field of urban dynamic simulation, scholars are committed to exploring different models and methods to improve simulation accuracy. However, while the overall accuracy is improved, the imbalance of data results in relatively low accuracy of minority samples. How to improve the simulation accuracy of minority samples has become an important issue. This paper will solve this problem at the sampling level. The experiment is based on the GlobeLand30 land use data in the three phases of 2000, 2010, and 2020, and uses the ProWSyn (Proximity Weighted Synthetic Oversampling Technique) algorithm to balance the data. Train the model with land use data from 2000 to 2010, verify the model with data from 2010 to 2020, and finally simulate the future based on data from 2020. The results show that the ProWSyn-MLP-CA model constructed in this paper has improved the accuracy of minority samples to varying degrees. In the case of using the same neural network model, the KAPPA value of a small number of land types increased by an average of 30% after ProWSyn equalization.

Keywords: ProWSyn, urban dynamic, data imbalance.

1. Introduction

As our country is in a period of rapid development, our lives are also undergoing rapid changes. The city and the surrounding environment have undergone major changes, so how to learn this trend of changes to make reasonable predictions for future urban development has become an important topic. Scholars have explored a wealth of methods in recent decades to realize "urban dynamic simulation".

Cellular Automata (CA) was proposed in the 1940s [1], and was first used to simulate urban expansion in the 1970s [2]. Because of its ability to calculate neighborhoods and combine spatial information, it has been widely used by scholars in the field of urban dynamic simulation. The composite model obtained through the combination of CA and neural network (Artificial Neural Network, ANN) [3] has attracted widespread attention since it was proposed. Since then, scholars have successively proposed CA coupled genetic algorithm [4], BP neural network [5], Markov model [6], nuclear learning machine [7], artificial immune system [8], particle swarm [9], which further improves the simulation accuracy of the model, and makes extraordinary contributions to this field.

In machine learning, data imbalance is a hot issue [10]. Data imbalance refers to the difference between the number of samples in the majority class and the sample in the minority class. Although the overall accuracy of the model is high, the accuracy of the minority class samples is relatively low.
Therefore, on the premise of ensuring the overall accuracy, how to improve the accuracy of minority samples has become a problem.

In the field of urban dynamic simulation, some scholars have studied this problem and used SMOTE [13], Decision Tree [14], Threshold Setting [15], Random Forest, and SMOTE-Tomek [16] to solve the problem. In this paper, the under-sampling ProWSyn algorithm [17] is selected to preprocess the experimental data set, and the processed data is input into the MLP-CA model for follow-up processing.

2. Research area and data processing

2.1. Research area overview
At the Chongqing Main City Metropolitan Area Symposium in May 2020, Chongqing announced that it will expand from the original 9 districts of the main city to 21 districts. The 21 districts are composed of the original 9 districts and 21 districts in western Chongqing, including Yuzhong and Dadukou, Jiangbei, Shapingba, Jiulongpo, South Bank, Beibei, Western Chongqing, Banan, Fuling, Changshou, etc.

![Figure 1. Location of Chongqing 21 District](image)

2.2. Data source and processing

2.2.1. Data Sources. This article will use the three phases of land use data for 2000, 2010, and 2020, which are derived from the global 30-meter resolution land cover data (GlobeLand30), of which the land use data for 2020 is the latest data on September 15, 2020. The data set divides land use types into seven categories: cultivated land, woodland, grassland, shrubland, wetland, water

The driving factor data used in the experiment includes DEM from the University of Maryland Geoscience Data Set (GLCF); roads (first-class roads, second-class roads, other roads), railways and river lines from OSM (Open Street Map) (Yangtze River, other tributaries).

2.2.2. Data processing. Due to the large area of Chongqing 21 District, the amount of dataset is big when the resolution is 30 meters, so this experiment decided to convert the 30-meter land resolution to 100-meter land resolution through ArcGIS resampling. The land classification maps for 2000, 2010 and 2020 are shown in the figure below.
3. Model construction and technical route

3.1. Model principle and construction

Since the neural network was proposed, it has attracted great attention from scholars. In the field of urban dynamic simulation, neural network is used to extract land conversion rules, so as to make reasonable predictions on the future distribution of land according to the land conversion rules. In this experiment, a common MLP model will be used as the basis for model prediction.

Cellular automata consists of three parts: land use conversion probability, neighborhood probability and random factors. In a grid with a range of 3*3, the type of the intermediate grid cell at the next moment will be determined by the state of its surrounding neighboring cells.

3.1.1. Land use transition probability. The deep neural network used in this article is mainly constructed using a python-based deep learning library. The first layer is the data input layer, with a total of 18 neurons, and each neuron corresponds to a spatial variable after data preprocessing; the middle layers is the hidden layer; the last layer is the output layer. The output layer uses the softmax excitation function to design neurons. There are 7 neurons in total, corresponding to the conversion probabilities of 7 land use types.
3.2. Principles of Sampling Technology
In the unbalanced data set, we call the class with more samples the majority class, and the class with fewer samples as the minority class. Classifiers tend to produce larger classification errors than minority samples. Many real-world problems are plagued by this phenomenon, such as medical diagnosis, information retrieval, detection of fraudulent calls, radar images, direct sales, and helicopter failure monitoring. Therefore, the identification of minority samples is crucial. There have been many attempts to solve the problem of unbalanced learning, such as various oversampling and under-sampling methods.

The under-sampling method removes some majority samples from an unbalanced data set, with the purpose of balancing the distribution between the majority samples and the minority samples, while the over-sampling method synthesizes the minority samples and adds them to the data set. Both under-sampling and over-sampling methods have been shown to improve the performance of the classifier on unbalanced data sets. When comparing over-sampling and under-sampling, under-sampling may remove important information from the original data, while oversampling does not encounter this problem.

3.2.1. The principle of ProWSyn. ProWSyn algorithm proposed by Sukarna Barua et al. is an oversampling technique of close-range weighted synthesis. It uses the distance information between the minority samples and the majority samples to assign weights to the minority samples.

The whole process is described as follows. ProWSyn finds the nearest K minority samples from each majority sample according to the Euclidean distance. The collection of all these minority samples forms the first partition P1 (close to level 1), the sample level closest to the boundary. Then, it finds the next K minority samples based on the distance from each majority class sample. Together these samples form the second partition, P2 (close to level 2). In this way, the samples are sequentially partitioned. The simulation partition is shown in the figure below. It can be seen that the minority samples are properly identified and divided according to their distance from the boundary, which also shows their importance in oversampling.

![Figure 5. ProWSyn algorithm schematic](image)

3.3. Technical route
The experiment in this article has three modules, namely module preprocessing, model training, module verification, and model prediction. In the data preprocessing module, the ProWSyn algorithm will be used to equalize the land use data; in the model training verification module, the balanced data will be input into the MLP-CA model for training, and after obtaining a better model, the simulation results are compared with the actual results; the trained model will be used to simulate the future in the final prediction module. The technical roadmap of this article is shown in the figure.
4. Experiment and result analysis

4.1. Data introduction

The training and testing data set of this article will use the land use status in 2010 as the label. There are 18 features, which are the distance to the highway, the railway, the first-level road, the second-level road, the other roads, and the distance from the Yangtze River, distance from other rivers in 2000, neighborhood of 7 types of land use in 2000, elevation, slope, roughness, land use change matrix.

The study area in this paper has a total of 2,876,961 data, and each data represents a grid. The distribution of the seven land use types is shown in the following table.

| Type             | Cultivated land | Woodland | Grassland | Shrubland | Wetland | Water | Construction land |
|------------------|-----------------|----------|-----------|-----------|---------|-------|-------------------|
| Num              | 2041010         | 510993   | 200065    | 34822     | 3682    | 9017  | 37372             |

The training data is obtained by stratified sampling from the data set, 30% of each land type is taken as training data, and the remaining 70% is used as test data. The following is the training data distribution table:

| Type             | Cultivated land | Woodland | Grassland | Shrubland | Wetland | Water | Construction land |
|------------------|-----------------|----------|-----------|-----------|---------|-------|-------------------|
| Num              | 609141          | 157418   | 55899     | 8854      | 965     | 13211 | 17597             |

4.2. Experiment procedure

This experiment will use 2000 data as the base year, 2010 as the final year, and simulate to the final year based on the base year data. We use above data to train the model; 2010 data as the base year and 2020
data as the final year to verification model; finally forecast the future land use situation based on the 2020 data. We use the original data set as the basic data for experiments. In the case of unbalanced data, we compare the simulated results with the real results, and calculate the overall kappa and the kappa values of each category. The data results are as follows:

**Table 3. Kappa in 2000-2010**

| Category          | Cultivated land | Woodland | Grassland | Shrubland | Wetland | Water | Construction land |
|-------------------|-----------------|----------|-----------|-----------|---------|-------|-------------------|
| Overall kappa     | 0.9703          | 0.8722   | 0.8349    | 0.4905    | 0.375   | 0.5732| 0.4826            |

From category kappa and overall kappa, the simulation results obtained through the model are generally consistent with the actual data. But the kappa accuracy of each land type is not consistent. Among them, the kappa accuracy of cultivated land, woodland, and grassland is higher, all better than 75%, but the kappa accuracy of wetland, construction land, shrubland, and Water is relatively low. Among them, the accuracy of wetland is the lowest, reaching 37.5%, and less than 40% belongs to lower accuracy range. The main reason for this situation is that the data is unbalanced. When the model is trained, it is easy to divide the minority class into the majority class, which will cause the kappa value of the majority class to be higher than the true result. In this way, if the minority samples are misclassified, the accuracy is relatively low.

4.2.1. **Model optimization.** In order to solve the above problems, this experiment decided to borrow the experience of previous scholars [16]. We use SMOTE algorithm to equalize the data to different land types. For each kind of balanced data, the same model was used to carry out experiments, and finally the better balanced data was selected by comparison. Next, we use ProWSyn to process the data to the optimal equilibrium level, and use the data to input into MLP-CA model and test the experimental results.

**Table 4. The training data set is balanced to different orders of magnitude**

|                    | Cultivated land | Woodland | Grassland | Shrubland | Wetland | Water | Construction land |
|--------------------|-----------------|----------|-----------|-----------|---------|-------|-------------------|
| raw data           | 609141          | 157418   | 55899     | 8854      | 965     | 13211 | 17597            |
| dataset 1          | 609141          | 157418   | 55899     | 8854      | 8854    | 13211 | 17597            |
| dataset 2          | 609141          | 157418   | 55899     | 13211     | 13211   | 13211 | 17597            |
| dataset 3          | 609141          | 157418   | 55899     | 17597     | 17597   | 17597 | 17597            |
| dataset 4          | 609141          | 157418   | 55899     | 55899     | 55899   | 55899 | 55899            |
| dataset 5          | 609141          | 157418   | 157418    | 157418    | 157418  | 157418| 157418           |
| dataset 6          | 609141          | 609141   | 609141    | 609141    | 609141  | 609141| 609141           |

**Table 5. Kappa values of different orders of magnitude**

|                    | Cultivated land | Woodland | Grassland | Shrubland | Wetland | Water | Construction land | overall |
|--------------------|-----------------|----------|-----------|-----------|---------|-------|-------------------|---------|
| raw data           | 0.9703          | 0.8722   | 0.8349    | 0.4905    | 0.375   | 0.5732| 0.4826            | 0.8746  |
| dataset 1          | 0.9935          | 0.9442   | 0.9485    | 0.5944    | 0.4136  | 0.6968| 0.6267            | 0.9358  |
| dataset 2          | 0.9890          | 0.9241   | 0.9172    | 0.5450    | 0.2310  | 0.7198| 0.5908            | 0.9208  |
| dataset 3          | 0.9882          | 0.9607   | 0.9820    | 0.6121    | 0.4940  | 0.7337| 0.6263            | 0.9494  |
| dataset 4          | 0.9771          | 0.9115   | 0.9134    | 0.4729    | 0.4043  | 0.7424| 0.5916            | 0.9105  |
| dataset 5          | 0.9488          | 0.8932   | 0.9394    | 0.4772    | 0.2359  | 0.7477| 0.6282            | 0.8969  |
| dataset 6          | 0.9511          | 0.8766   | 0.9926    | 0.5902    | 0.6131  | 0.7700| 0.6225            | 0.8418  |

As shown in the above table, with the number of minority samples gradually balanced to the majority samples, the overall kappa value increased at first, reached the highest value in dataset 3, and then
showed a downward trend. In a few land types, the best precision performance of wetland and water is dataset 6, shrubland is the best performance in dataset 3, and construction land performance the best in dataset 1. Among them, the accuracy of shrubland and construction land in data set 6 is similar to its best accuracy, and the accuracy of most land types is reduced. Because the accuracy of most classes has been reduced, the overall accuracy has also been reduced accordingly, so after accuracy comparison, this experiment will select data set 6 as the optimal balanced data set.

Import the balanced training data into the MLP-CA model with the same structure and calculate the kappa value. The results are shown in the following table:

| Category | Cultivated land | Woodland | Grassland | Shrubland | Wetland | Water | Construction land |
|----------|----------------|----------|-----------|-----------|---------|-------|-------------------|
| Kappa    | 0.9383         | 0.8321   | 0.9862    | 0.6445    | 0.4572  | 0.8185 | 0.6295            |
| Overall Kappa |           |          |           |           |         |       | 0.8673            |

After the data is equalized by the ProWSyn algorithm, it can be found from the results of the model that the accuracy of the minority samples has been improved to varying degrees. Among them, the kappa value of the water increased by 42%, the shrubland increased by 30%, and the construction land increased 29%, wetland increased by 21%. In most categories, the accuracy of cultivated land and woodland has decreased, and the accuracy of grassland has increased.

From the experimental results, ProWSyn can effectively improve the simulation accuracy of a small number of samples, and achieve the purpose of this experiment. Next, we will verify the generalization ability of the model through 2010 data simulation to 2020.

Table 6. Simulated and actual confusion matrix from 2000 to 2010

From the simulation results, the simulation results and the actual results are generally not much different, especially in urban land, grassland, woodland and other land types, the simulation accuracy is higher, roughly in line with the actual situation of land use in 2020. From the point of view of the kappa value of each type of land, the kappa value of grassland, cultivated land and shrub land is more than 75%. In a few samples, the kappa value of woodland, water and construction land is higher, which indicates that the model used in this experiment can accurately simulate the real situation in 2020.

Table 7. Simulated 2020 land use data kappa value of each category
4.2.2. Model comparison. In order to verify the feasibility of the ProWSyn-MLP-CA model proposed in this paper, the coupling model constructed by selecting 6 equalization algorithms is compared with the model in this paper. The comparative experiment will use the 2010 land use data to predict the 2020 land use data, and calculate the kappa value of each category and overall under different models, as shown in the following table:

Table 8. Experimental results of coupling models of different equalization algorithms

| Equalization Algorithm | Cultivated land | Woodland | Grassland | Shrubland | Wetland | Water | Construction land | Overall  |
|------------------------|----------------|----------|-----------|-----------|---------|-------|-------------------|---------|
| ProWSyn-MLP-CA         | 0.8753         | 0.7946   | 0.9081    | 0.7322    | 0.4324  | 0.7738 | 0.9178           | 0.8544  |
| Distance_SMOTE-MLP-CA  | 0.8479         | 0.7562   | 0.9141    | 0.6681    | 0.3734  | 0.5575 | 0.8425           | 0.8142  |
| LLE_SMOTE-MLP-CA       | 0.9224         | 0.7950   | 0.7604    | 0.6057    | 0.2208  | 0.6596 | 0.8662           | 0.8413  |
| SMMO-MLP-CA            | 0.8390         | 0.6883   | 0.5383    | 0.5785    | 0.1818  | 0.5395 | 0.7433           | 0.7295  |
| polynom_fit_SMOTE-MLP-CA | 0.8372    | 0.6784   | 0.5763    | 0.4323    | 0.0218  | 0.7274 | 0.8471           | 0.7462  |
| safe_level_smote-MLP-CA | 0.9102   | 0.7721   | 0.6201    | 0.3618    | 0.0639  | 0.6549 | 0.7471           | 0.7911  |
| DE_oversampling-MLP-CA | 0.8946         | 0.7916   | 0.7951    | 0.4140    | 0.2030  | 0.6457 | 0.8658           | 0.8237  |

As can be seen from the table above, LLE_SMOTE-MLP-CA has the highest kappa value for cultivated land and woodland, and Distance_SMOTE-MLP-CA has the highest kappa value for grassland and wetland. ProWSyn-MLP-CA proposed in this paper has the highest kappa value on shrubland, water, and construction land. The ProWSyn-MLP-CA has the highest overall kappa value. From the experimental results, the ProWSyn-MLP-CA model proposed in this paper improves the kappa value of the majority and minority samples at the same time, which is in line with the expectations of this experiment.

4.3. Future simulation

Taking 2020 as the base year, the data of neighborhood factors and driving factors in 2020 are obtained to simulate the land use in 2030. Firstly, the Markov module integrated in IDRISI is used to obtain the land transfer probability in 2020. As shown in the table:

Table 9. The number of grids in each year

| Year | Cultivated land | Woodland | Grassland | Shrubland | Wetland | Water | Construction land |
|------|----------------|----------|-----------|-----------|---------|-------|-------------------|
| 2000 | 1827425        | 472255   | 167699    | 26564     | 2898    | 39634 | 52794             |
| 2010 | 2030472        | 524727   | 186332    | 29515     | 3219    | 44037 | 58659             |
| 2020 | 1846092        | 508566   | 217671    | 45534     | 950     | 64232 | 194116            |
| 2030 | 1659044        | 515452   | 266449    | 61441     | 830     | 88184 | 285559            |
The simulation results in 2030 are shown in the figure. From the table and figure, we can find that most land types have increased, and the construction land has increased by approximately 47%, indicating that in the next ten years, the scope of the city in Chongqing’s 21 districts will be further expanded; cultivated land will increase slightly compared with 2020, shrubland, woodland, and water bodies will increase more in 2030, while the area of grassland will be larger. At the same time, the area of wetland will continue to decrease.

5. Conclusion
In this paper, Chongqing 21 district is taken as the research area. The land use data of 2000, 2010 and 2020 are used as data support, and 18 data such as elevation, slope, and distance from river, first-class road and second-class road are used as driving factors. In this experiment, SMOTE algorithm is used to balance the training data from 2000 to 2010 to different orders of magnitude, and the optimal data level is found according to the experimental results; then, the training data is equalized to the optimal data level by using ProWSyn equalization algorithm as the training data of MLP-CA model; then the model is verified with 2010 data as the base year and 2020 data as the final year, and is compared with six different equalization calculations. Finally, based on the land use data in 2020 and driving factors, the land use data in 2030 is predicted.

1) In this paper, the SMOTE algorithm is used to balance the data to different levels to find the optimal level of balanced data, so that the training data can be balanced to the level of optimal balanced data in the equalization step.

2) ProWSyn oversampling algorithm is used to solve the problem of data imbalance. After the training data is balanced to the optimal equilibrium data level by ProWSyn algorithm, the MLP-CA model is trained to extract the conversion rules of seven land-use types, and then the conversion rules are multiplied with neighborhood and random factors to obtain the overall conversion probability of seven land-use types. Finally, the simulated land use data are obtained.

References
[1] UlamS. Adventuers of a Mathematican (Charles Scribnerssions New York), 1976
[2] Tobler W R. A computer movie simulating urban growth in the Detroit region [J] Economic Geography, 1970
[3] Li Xia, YE Jia-an. Cellular automata based on neural Network and Simulation of complex land use System [J]. Geographical research, 2005 (01): 19-27.

[4] Yang Qingsheng, Li Xia. Automatic Acquisition of CA model parameters based on Genetic Algorithm -- A case study of Urban development simulation in Dongguan city [J]. Geographical research, 2007 (02): 229-237.

[5] Li Mingyue, Lai Xiaojuan. Evaluation of Urban Land Ecological Security based on BP Neural Network -- A Case study of Guangzhou city [J]. Economic geography, 2011, 31

[6] Hou Xiyong, Chang Bin, YU Xinfang. Research on land use change in Hexi Corridor based on Ca-Markov [J]. Journal of Agricultural Engineering, 2004 (05): 286-291.

[7] Xiaoping Liu, Xia Li, Xun Shi, Shaokun Wu, Tao Liu. Simulating complex urban development using kernel-based non-linear cellular automata [J]. Ecological Modelling, 2007, 211 (1).

[8] Xiang Zhao, Xiaoya Ma, Wenwu Tang, Dianfeng Liu. An adaptive agent-based optimization model for spatial planning: A case study of Anyue County, China [J]. Sustainable Cities and Society, 2019, 51.

[9] Ma Shifa, HE Jianhua, Yu Yan. Spatial Optimization model of urban Land Use based on ParticleSwarm Optimization [J]. Journal of Agricultural Engineering, 2010, 26

[10] Wei Xun. Study on integrated Classification method of Category imbalance and cost sensitive Data [D]. University of Science and Technology of China, 2017.

[11] Wu Yuqian, WANG Junli, Yang Li, YU Miaomiao. A review of cost sensitive deep learning methods [J]. Computer Science, 2019, 46 (05): 1-12.

[12] XIONG B Y, WANG G Y, DENG W B, et al. Unsaturated data undersampling method based on sample weight [J]. Journal of Computer Research and Development, 2016, 53 (11): 2613-2622. (in Chinese)

[13] Feng Yuan. Construction of integrated model based on Multi-layer perceptron and CA and Land use scenario simulation [D]. Chongqing: Chongqing University of Posts and Telecommunications, 2019.

[14] Yang Ernu. Land use classification of Xishan Coalfield based on decision tree [J]. Land resources of North China 2018(02):103-105.,

[15] Ou Zhaokeng. Construction of ANN-CA Model under Markov Constraint and Land use Simulation study in Minhou County [D]. Fujian Agriculture and Forestry University.

[16] Lei Yuhang, LIU Minghao, Chen Sinan, et al. Comparative Analysis of Data Equalization Algorithm in the Application of Spatio-temporal dynamic Model -- A Case study of the Three Gorges Reservoir area [EB/OL]. Beijing: China Science and Technology Papers online [2019-12-25].

[17] Barua S., Islam M.M., Murase K. (2013) ProWSyn: Proximity Weighted Synthetic Oversampling Technique for Imbalanced Data Set Learning. In: Pei J., Tseng V.S., Cao L., Motoda H., Xu G. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2013. Lecture Notes in Computer Science, vol 7819. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-37456-2_27