Simulating urban growth through case-based reasoning

X. Ye, W. W. Yu, W. H. Yu, and L. N. Lv

Heilongjiang University of Science and Technology, Harbin, Heilongjiang, China; Hebei Institute of Investigation & Design of Water Conservancy & Hydropower Co., Ltd, Shijiazhuang, Hebei, China

ABSTRACT
Case-based reasoning (CBR) simplifies knowledge acquisition and is suitable for researching complex geographical problems. However, CBR analyses of land-use changes are difficult to apply in the study of urban growth due to shortcomings in the case structure and model algorithms. In response, this study proposes a three-step urban-growth simulation model based on CBR (UGSCBR). First, to adapt the CBR to an urban-growth simulation process, the characteristics of regional differentiation in geographical spaces are determined. Second, a comprehensive retrieval method is developed that improves upon traditional case-retrieval methods by giving full play to the comprehensive function of each component of the case. Third, a quantity demand constraint indirectly adds a time factor to solve the initial blurriness of the traditional CBR-inference cycle. Taking Jixi city as the research area, we test the accuracy of the proposed model. The total accuracy of simulation results is 95.4%, and the Kappa is 87.4%. The figure of merit and Mathews correlation coefficient are 0.151 and 0.23, respectively, indicating that the model can meet the application requirements. The results show that the UGSCBR model has strong flexibility and simplicity, and it provides an effective prediction method for urban growth.

Introduction
Urban growth is the process of cities spreading outward to occupy other types of land (Mookherjee & Hoerauf, 2015). It is the result of economic, social, political, legal, institutional, and environmental factors that interact in a parallel and sometimes conflicting manner (Campos et al., 2018). Spatiotemporal land-use and land-cover change simulations are effective and reproducible tools for assessing and analyzing the process of dynamic urban change (Liu et al., 2017). Scholars have used a variety of simulation models to study the expansion of different cities. For example, the Markov model (Ahmad et al., 2017; Coppel, 2007) can predict changes in quantity over a certain period time, but it is a non-spatial model. The system dynamics model (Z. Liu et al., 2019; Tan et al., 2018; Wu & Ning, 2018) can effectively explain the function of complex systems and their relationship with structure, but the spatial scale effect in earth science cannot be integrated into this model. The Bayesian is an effective method for simulating land-use change in spatial regions, but it requires extensive assumptions, such as a complete dataset, no preferred selection, and discontinuous variables network (Ouyang et al., 2019). Cellular automata (CA; Cheng & Masser, 2004; Firozjaee et al., 2019; Jafari et al., 2017; Shafizadeh-Moghadam et al., 2021) is a very useful tool for simulating complex systems and is a common method of simulating urban growth; however, the acquisition process of its transformation rules is complex, and it is difficult to describe the geographical significance of complex rules. From the perspective of existing applications, these methods and models each have their own pros and cons (Du et al., 2010), making it necessary to explore new ideas that can solve the problems of urban growth simulation.

Case-based reasoning (CBR) is an important approach in the field of artificial intelligence that relies on having enough historical data to conduct quantitative analyses and predictions of phenomena without knowing the development mechanisms (Du et al., 2010; Mantaras et al., 2006). It is essentially memory-based reasoning, using specific case knowledge from previous experiences to solve similar problems; it therefore conforms to human cognitive processes (Aamodt & Plaza, 1994; Watson, 1999). It eliminates the need to display domain knowledge models, avoids the constrictions of knowledge acquisition, and has the advantages of simplifying knowledge acquisition and knowledge accumulation while improving problem solving efficiency and solution quality (Du et al., 2002).

Currently, in the field of GIScience, CBR is being applied to forest fire monitoring (W. Liu et al., 2015), urban disaster prevention and mitigation (Yu & Li, 2018; Yu et al., 2018), water environment simulation (Liao et al., 2019), mapping (Dou et al., 2015; Du et al., 2013; Machado et al., 2019), and other research directions. Few studies, however, have been conducted on the spatial patterns and evolution of geographical phenomena using CBR. At present, such research mainly

CONTACT L. N. Lv; lvlinawode@163.com; Heilongjiang University of Science and Technology, Harbin 150022 China
© 2022 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
includes two types of studies: a CA simulation process of land-use change constructed by replacing static rules with CBR, and a CBR reasoning that integrates CBR and land-use change. The former is represented by the CBR-based CA model proposed by Li and Liu (2006), which showed that CBR can effectively replace static rules and solve the problems regarding the internal differences of complex systems that the traditional rule-based CA model could not. However, this research focused on the mining of transformation rules using CBR, without in-depth discussions on CBR’s application in land-use change analysis. Du et al. (2010) proposed using CBR to analyze land-use change and showed that the CBR method is simple and more flexible than the CA model (Du et al., 2012). However, based on the present literature, it is difficult to simulate the spatial pattern formation processes of a complex system, and the single-direction reasoning of previous case-retrieval results lacks comprehensive analysis based on historical experience. In addition, the lack of key ideas on the time factor control and the blurry CBR-reasoning cycle makes it difficult to determine the simulation result period.

In summary, the CBR method can implicitly describe the rules of urban pattern transformations based on historical cases. It breaks through the constrictions of traditional models’ rule acquisitions and has a better effect in terms of simulating complex areas. However, the existing research employing CBR to study land-use change has shortcomings with respect to case structures and model algorithms, making the CBR reasoning incomplete. In addition, the differences between the research objectives of those studying land-use change and urban growth make it difficult to apply this method to land-urbanization research. Thus, this study aims to improve the existing CBR land-use change analysis method (Du et al., 2010) and establish a CBR urban-growth simulation model to analyze urban spatial pattern evolution. The following problems must be solved within the modeling process:

(1) Improve the existing representation modes for land-use change cases by proposing a set of case structures that adapt to the urban-growth simulation process.
(2) Develop a comprehensive retrieval approach to solve the problem of single reasoning in traditional research.
(3) Study the introduction of a time factor to solve the problem of blurry CBR-inference cycles.

Accordingly, this study proposes an urban-growth simulation model based on the CBR approach (UGSCBR) and applies it to an urban-growth simulation of Jixi city in China from 2005 to 2015.

The remainder of this article is organized as follows. Section 2 describes the proposed UGSCBR model in detail. Section 3 describes the study region, lists the data, and provides the experimental results. Section 4 discusses the results, and Section 5 concludes.

**The urban-growth simulation case-based reasoning approach**

The basic idea of the UGSCBR model is as follows (Figure 1): Taking the complete description of each land unit in time and space, the land-evolution type is inferred. The “new case”, that is the urban-growth result, is then obtained by retrieving the most similar previously known case. To address the three key problems mentioned in the introduction, the UGSCBR model includes three steps: “case organization”, “case retrieval”, and “case constraint”. The model retrieves known cases that are similar to the unknown case and deduces the results of the unknown case using these levels. Figure 2 provides a detailed illustration of the modeling process.

**Case organization**

The case represents the basic unit and essence of CBR (Holt, 1999), as its structure and representation mode determine how CBR works. The existing study (Du et al., 2010) on CBR land-use change takes land parcels

![Figure 1. The basic idea of the UGSCBR model](image-url)
as the case unit, making it difficult to simulate the spatial-pattern formation processes of a complex system. Moreover, the case representation structure of land-use change cannot be fully adapted to the simulation of urban growth. Therefore, it is necessary to develop a set of case structures that can better express the change process of complex spatial patterns and that can adapt to the research objectives of urban-growth scholarship.

Regular, discrete raster pixels are the most commonly used mode in the study of geographical problems. This representation mode can effectively simulate the formation process behind the spatial patterns of complex systems by generating global patterns through local, individual behaviors. Therefore, the UGSCBR model proposed in this study is represented using this mode.

This study proposes a case representation structure (initial state, geographic features, and result) based on the existing three-component mode of land-use change (Du et al., 2010). The meaning of each component is as follows: (1) the initial state describes the land-use type at the beginning of the unit, (2) the geographic features include a set of spatial data indices that affect urban growth, and (3) the result involves whether the unit is ultimately transformed into urban land. The mode can be regarded as the process of land-unit (case) change within a certain period. In other words, the land’s initial state is affected by its geographical features, which will help achieve the end result. Cases can be defined by the following equation:

\[ \text{CASE}_i = LU_i, \text{Index}_{i1}, \text{Index}_{i2}, \ldots, \text{Index}_{in}, \text{Result}_i \] (1)

where \( i \) is the case number, \( LU_i \) is the case’s initial state, and \( \text{Result}_i \) is the result of case \( i \), which is a Boolean variable with “1” indicating urbanization and “0” indicating non-urbanization. \( \text{Index}_{1i}, \text{Index}_{2i}, \ldots, \text{Index}_{ni} \) represent the multidimensional geographical index of case \( i \), with a total of \( n \). For example, in a case describing the change process of a land unit from 2005 to 2010 (Figure 3), the “initial state” (\( LU \)) is the land-use type of the unit in 2005, the “geographic features” (\( \text{Index}_{i} \)) are the geographic-feature indices of the unit in 2005, and the “result” is the urban-growth result of the unit in 2010.

This model improves the adaptability of the CBR model’s case structure to the study of urban growth by taking into account the effect of land-use type (initial state) and spatial environment (geographical features). It also simplifies the description of the “results” to urbanization and non-urbanization.

The CBR model is a process of inferring the results of a new period from a past one. To distinguish the role of cases in different periods, cases expressing historical changes in urban
growth are regarded as “geographical cases”, while those used to simulate urban growth are considered “simulated cases”. By collecting grid cells in space, a “geographic case database” (GCDB) and “simulation case database” (SCDB) can each be constructed. Table 1 further illustrates the terms proposed in the UGSCBR model.

**Case retrieval**

The purpose of case retrieval is to find the most similar cases. We quantitatively calculate the similarity coefficient between cases and then deduce the simulated case result. The simulated case is deduced by finding the geographic case with the shortest distance in attribute space from the simulated case. The basic process of UGSCBR case retrieval is as follows: Each experience case is searched, and the most similar geographical case is taken as the result of the simulated case. To give full play to the comprehensive function of each component of the case, this study proposes a comprehensive retrieval strategy, which uses “initial state” to induce, “geographical features” to retrieve, and “results” to reason the case.

Different land-use types have different urbanization conditions. To reflect the role of the initial state, this study proposes the idea of induction and then retrieval; that is, all cases are summarized and grouped according to their initial state, so that each group of simulated cases can only be retrieved from geographical cases in the same group. For example, a simulated case with an initial state of arable land only needs to be retrieved and calculated with the geographical cases in the GCDB that also have arable land as their initial state.

Case retrieval is realized through similarity calculation of the cases’ geographical features, with the nearest neighbor method used most often. The smaller the coefficient value calculated by this method, the higher the similarity between cases. The following equation describes this similarity:

\[
SIM = \sqrt{\sum_{i=1}^{n} w_i^2(p_i - q_i)^2}
\] (2)
In this equation, \( SIM \) represents the case’s similarity coefficient value, \( i \) is the index number, \( p_i \) is the \( i \) of the geographic case, \( q_i \) is the \( i \) of the simulated case, \( n \) is the index quantity, and \( w_i \) is the weight assigned to \( i \). While there are many ways to determine the weight, referring to relevant research (Li & Liu, 2006), the entropy weight method is adopted in this study. Before determining the weight of each index, the variables need to be normalized so that their values fall between \([1,0]\). This can be calculated as follows:

\[
H_n = \sum_{i=1}^{n} a_i \log(1/a_i) / \log(n) \tag{3}
\]

\[
a_i = \frac{p_i}{\sum_{i=1}^{n} p_i} \tag{4}
\]

where \( H \) is entropy and \( n \) is the total sample number.

The entropy falls within the range of \([0,1]\). The smallest value of 0 represents the maximum amount of information exhibited in the variable, and the largest value of 1 indicates the minimum amount of information. Thus, the amount of information is proportional to the form: \( 1-H_n \). A feature with more information is expected to have greater weight. Thus, the entropy weight of the \( i \)th indicator can be expressed as:

\[
w_i = \frac{1 - H_{ni}}{n - \sum_{i=1}^{n} H_{ni}} \tag{5}
\]

The similarity coefficient between each simulated case and all geographical cases can be calculated by Eq. (2) to match the most similar geographical case. To give full play to the outcomes of urbanization and non-urbanization, this study proposes a retrieval method based on retrieval measurements, that is, a simultaneous retrieval of the most similar geographical cases from the two kinds of “results”. The retrieval process for a certain simulated case \( j \) is as follows (Figure 4):

1. Calculate the similarity coefficient \( SIM \) between the simulated case and each geographic case with the same group (the same “initial state”) using Eq. (2). The calculated \( SIM \) number is the same as the number of geographic cases in the group.

2. Record the similarity coefficient values of the most similar geographic cases in urbanized (result is 1) and non-urbanized (result is 0) locations as \( SIM_{min-1} \) and \( SIM_{min-0} \). The minimum value \( (SIM_{min}) \) is defined as the “retrieval measure” of simulated case \( j \).

3. The result of simulated case \( j \) is deduced by a retrieval measure, and the reasoning formula is as follows:

\[
Result_j = \begin{cases}
0, & SIM_{min} = SIM_{min-1} \\
1, & SIM_{min} = SIM_{min-0}
\end{cases} \tag{6}
\]

Applying the above method to the retrieval process of each simulated case, the results of all the simulated cases can be obtained.

**Case constraint**

Within the actual geographical environment, different land-use types tend to have specific spatio-temporal change characteristics, and the similarities between these characteristics in the time domain are suitable for solving the issues with the case retrieval methods (Qian et al., 2007). To determine the reason for the case simulation’s result, it is necessary to control the
case change process in the time domain, that is, the period of case change guided by a time factor. However, given the lack of a time factor control in traditional CBR-reasoning processes (Du et al., 2010) caused by employing a blurry CBR-reasoning cycle, the period of the simulation results is difficult to determine.

In the UGSCBR model, although it is possible to simulate all cases through a complete case retrieval, it is not possible to determine the number of cases that have changed during the simulation period. This study proposes a method to control the number of case changes after case retrieval by calculating the quantity of urban growth (quantity demand [QD]) in the forecast year, introducing the time factor in an indirect way to solve the above problem. First, we reclassify the simulated case based on the retrieval measure. When $SIM_{min} = SIM_{min-1}$, we add the case to the quasi-urbanized case set, and when $SIM_{min} = SIM_{min-0}$, we add the case to the quasi-non-urbanized case set. Then, Markov (Blanchard et al., 2010), spatial superposition, or other methods can be used to calculate and compare the urban growth (QD, in which i is the type of initial state) and number of cases ($NC_i$) in the quasi-urbanized case set. The reasoning behind the three possible scenarios is as follows (Figure 5):

1. When $QD_i = NC_i$, it infers that all cases in the quasi-urbanized case set of type i are transformed into urban land, while cases in the quasi-non-urbanized case set are not transformed. This situation has a low occurrence probability.

2. When $QD_i < NC_i$, the cases in the quasi-urbanized case set are sorted from small to large according to the $SIM_{min}$. This infers that $QD_i$ cases are transformed into urban land, while other cases in the set and in the quasi-non-urbanized case set are not transformed.

3. When $QD_i > NC_i$, it infers that all cases in the quasi-urbanized case set of type i are transformed into urban land, and the cases in the quasi-non-urbanized case set are sorted from small to large through $SIM_{min-1}$. To ensure there are enough cases, such cases are transformed into urban land to meet the needs of urban growth, while other cases are not transformed.

**Case study**

**Study area and data**

The proposed UGSCBR model was applied to the simulation of urban growth in the main urban area of Jixi city. Jixi is a southeast city in Heilongjiang Province, China (Figure 6). It has jurisdiction over nine county-level administrative regions, covering a total area of approximately 22,550 km², and has a total population of 1.72 million. This region has a cold temperate continental monsoon climate, and the terrain mainly includes mountains, hills, and plains. By 2018, its economic output was $7.7 billion. The spatial distribution of coal resources and geological conditions affects urban development in this region. Its urban spatial structure is dispersed, and the degree of aggregation is low. The built-up urban area only accounts for 37% of the total urban land area. As a mining resource-based city with strong resource support capacity, the study area continues to expand, and its evolution process is complicated. As such, urban development regulation is not obvious, and it is difficult to determine a set of ideal rules for urban space transformation. Thus, Jixi city serves as an ideal location to test the effectiveness of the UGSCBR method.
We obtained land-use data for 2005, 2010, and 2015 from the classifications of the Thematic Mapper satellite images (pixel size 30 m × 30 m). The data were divided into six types of land use (urban land, arable land, woodland, grassland, water, and unused land), and CGC2000 was taken as a unified coordinate system. Referring to relevant literature (Aburas et al., 2017; Chen et al., 2016; Liang et al., 2018b), we standardized and divided various land-use drivers into distance and terrain factors with the same resolution as the land-use data. Table 2 lists the data from this study. This study used ArcGIS10.2 as the data processing and model operation software.

**Table 2. Research data and sources.**

| Category          | Data                      | Year       | Data resource         |
|-------------------|---------------------------|------------|-----------------------|
| Land use          | Land-use data             | 2005–2015  | Image interpretation  |
| Terrain           | DEM                       | 2010       | http://www.gscloud.cn |
| Distance index    | Distance to road ($D_{road}$) | 2005–2010  | Using the “Euclidean distance” function to obtain the distance of all grid cells to the nearest railway or highway. |
|                   | Distance to city center ($D_{city}$) |          | Using the “Euclidean distance” function to obtain the distance between all cells and the city center. |
|                   | Distance to city edge ($D_{edge}$) |          | Using the “Euclidean distance” function to obtain the distance of all grid cells to the nearest urban land. |
|                   | Distance to water ($D_{water}$) |          | Using the “Euclidean distance” function to obtain the distance of all grid cells to the nearest body of water. |

**Model implementation**

This study simulated the research area’s urban growth in 2015. The experiment established the SCDB between 2010 – 2015 with a transition period of 5 years, based on the 2010 land-use condition data, and established a GCDB of historical cases from 2005 – 2010. According to the UGSCBR model’s case structure, the simulated case is expressed as follows: the initial state describes the land-use type of the unit in 2010, the geographic features are the various indices in 2010 (Table 2), and the result is the urban-growth simulation result in 2015, which took a null value when building the database. For the geographic case, the initial state reports the land-use type of the unit in 2005, the geographic features are the various indices in 2005 (Table 2), and the result describes whether the unit had been transformed into urban land in 2010. Geographic cases need to be grouped according to the initial state. We did not consider changes to water and urban land, and only constructed the GCDB with
Table 3. GCDB sampling numbers (cells).

| Initial state    | Arable land | Woodland | Grassland | Water | Unused land |
|------------------|-------------|----------|-----------|-------|-------------|
| urbanized case   | 10,156      | 986      | 55        | 0     | 280         |
| Non-urbanized    | 10,033      | 10,180   | 3229      | 0     | 3540        |
| Total            | 20,189      | 11,166   | 3284      | 0     | 3820        |

Table 4. The GCDB for 2005–2010.

| ID   | LU | DEM D<sub>l</sub> | D<sub>r</sub> | D<sub>max</sub> | D<sub>water</sub> | D<sub>edge</sub> | Result |
|------|----|------------------|--------------|----------------|------------------|----------------|--------|
| 0    | 1  | 0.323            | 0.384        | 0.428          | 0.583            | 0.334          | 1      |
| 1    | 2  | 0.051            | 0.760        | 0.144          | 0.051            | 0.038          | 1      |
| 2    | 3  | 0.244            | 0.075        | 0.119          | 0.024            | 0.077          | 1      |
| 3    | 4  | 0.543            | 0.536        | 0.926          | 0.217            | 0.070          | 1      |
| 4    | 5  | 0.545            | 0.983        | 0.666          | 0.254            | 0.084          | 0      |

urban areas for 2015 and the actual situations derived from land-use data. The results show that the actual situations (Figure 7) were similar to the simulated ones (Figure 7). A confusion matrix and error map of the concordance between the simulated and actual situations was further obtained to conduct a quantitative analysis (Table 7 and Figure 8). This matrix was calculated based on a cell-on-cell spatial overlay of these two situations (total simulation accuracy: 95.4%). The simulation accuracy of the urbanized and non-urbanized cases was 90.4% and 97.0%, respectively. In addition, the kappa coefficient was 87.4%. Overall, the results show that the UGSCBR model has a high simulation accuracy.

The proportion of unchanged regions in the whole study area is often high in urban simulations. The total accuracy of the confusion matrix and kappa coefficient involves these unchanged parts in the calculation range, which leads to a high calculation value. Employing the “figure of merit” (FoM) indicator can help avoid this situation (Chen et al., 2013). The FoM index is better than the kappa in evaluating the accuracy of simulated changes and can be calculated by Eq. (7; Pontius et al., 2004). In addition, the Matthews correlation coefficient (MCC) is considered as a balance measure for evaluating the quality of binary classifiers or models and can be calculated by Eq. (8; Boughorbel et al., 2017; Kantakumar et al., 2019).

In this equation, Misses is the area of non-urban land that is transformed into urban land in the actual scenario but not in the simulation scenario. Hits is the area of non-urban land transformed into urban land in both scenarios (the correctly transformed area). False alarms is the area of non-

\[ \text{Matthews correlation coefficient (MCC)} = \frac{\text{Hits}}{\sqrt{\text{Hits} + \text{False alarms}} \times \text{Correct rejections} - \text{False alarms} \times \text{Misses}} \]

\[ \text{Figure of merit} = \frac{\text{Hits}}{\text{Misses} + \text{Hits} + \text{False alarms}} \]

Table 5. Case constraint scenarios (cells).

| Initial state | Arable land | Woodland | Grassland | Water | Unused land |
|---------------|-------------|----------|-----------|-------|-------------|
| QD            | 9670        | 892      | 52        | 0     | 201         |
| NC            | 26271       | 3158     | 18        | 0     | 660         |

Table 6. Index weights for each group of geographic cases.

| Drivers | DEM | D<sub>l</sub> | D<sub>r</sub> | D<sub>max</sub> | D<sub>water</sub> | D<sub>edge</sub> |
|---------|-----|--------------|--------------|----------------|------------------|----------------|
| Arable  | 0.0890 | 0.3468 | 0.0638 | 0.1805 | 0.3199 |
| Woodland| 0.0823 | 0.3510 | 0.0602 | 0.2026 | 0.3038 |
| Grassland| 0.0994 | 0.1895 | 0.0985 | 0.2595 | 0.3531 |
| Unused  | 0.1434 | 0.3072 | 0.0542 | 0.2769 | 0.2183 |

The UGSCBR must be validated when it is applied to the simulation of actual cities. The simplest validation method is to visually compare the simulated patterns with the actual ones (Liu et al., 2013). Therefore, visual inspections were conducted to compare the simulated arable land, woodland, grassland, and unused land. Each group of GCDB was built by randomly sampling 10,000 urbanized and non-urbanized cases, respectively, and collecting all cases in this group if their numbers were insufficient. Table 3 shows the statistics of each case group, and Table 4 presents the GCDB.
urban land that is not transformed into urban land in the actual scenario but is in the simulation scenario. Correct rejections is the area of non-urban land that is unchanged in both scenarios.

The FoM index and MCC were calculated as 0.151 and 0.23, reconfirming the UGSCBR model’s high simulation accuracy and ability to meet the application requirements.

Discussion

Influence of geographical case number on accuracy

The simulation accuracy of each land type’s urbanization was further analyzed, and the FoM index of the arable land, woodland, grassland, and unused land simulations was calculated as 0.156, 0.129, 0.010, and 0.055, respectively. The results show that the simulation accuracy for the different land types varies greatly, and the key reason for this difference is the varying number of recorded historical experiences for each case type. After comparing the FoM with the number of cases (Figure 9), we found a significant positive relationship between the FoM index and the total number of geographic cases in each group. In addition, the shortage of geographic cases affected the simulation accuracy of grassland and unused land as well as the overall accuracy. Although the non-urbanized geographic cases of arable land and woodland were the same, the differences between the urbanized geographic cases caused the low simulation accuracy of the woodland. Therefore, increasing geographic cases plays a positive role in improving simulation accuracy, and the number of geographic cases with different results in each initial state will affect the simulation results.

Influence of case constraints on accuracy

The UGSCBR model’s constraint process consists of two parts. On the one hand, the similarity coefficients ($SIM_{min-t}$ and $SIM_{min-o}$) of the urbanized and non-
urbanized geographic cases that were most similar to the simulated cases were comprehensively retrieved, and the $SIM_{min.0}$ was used to constrain the transformation sequence of the simulated cases. On the other hand, the quantity demand ($QD_i$) was used to control the number of cases transformed into urban land. The two constraint methods are tested below.

We can analyze the simulated case accuracy, as affected by the constraint, to study the process of the $SIM_{min.0}$ constraint. As shown in Table 8, we extracted the cases that could not be transformed into urban land due to the constraint during the simulation process: in 2015, 284 cases were transformed and 1162 cases were not, under the actual scenario. We then extracted the supplementary cases that were transformed into urban land due to the constraint, among which 306 cases were transformed and 1140 cases were not, under the actual scenario. In the cases affected by the constraint, their

![Figure 8. Error map.](image)

| Constraint cases | Supplement cases |
|------------------|------------------|
| Correct          | 284              |
| Incorrect        | 1162             |
| Accuracy         | 19.6%            |
|                  | 1140             |
|                  | 21.2%            |

Table 8. The constraint effect on the result by $SIM_{min.0}$.
However, simulated the most were dropped. Therefore, the quantity demand method was proven to be effective.

### Influence of other factors on accuracy

The research area observed in this study concerns a mining resource-based city. Compared with ordinary cities, the urban land is more scattered (Duncan et al., 2009; Schueler et al., 2011; Zhang et al., 2017) and the laws of urban development are not obvious. New coal mines, industrial parks, and changes in transformation policies will lead to unpredictable development of the city (Ye et al., 2021). The results of this study are compared to that of similar studies that use multiple models to simulate the urban growth of Tehran, Mumbai, and Pune (Kantakumar et al., 2019; Shafizadeh-Mohgadam et al., 2017, 2015), and the FoM is above 0.40. According to the distribution and shape of urban land, most of Tehran’s urban land is concentrated in the north, with an overall high degree of urban concentration, and the southern part of the city is the main direction of urban growth; Mumbai’s urban-land connectivity is relatively high, and urban growth is mainly based on internal filling. However, the urban land in Jixi city is mainly scattered in the south, middle, and east of the study area, with more directions and more complicated urban growth patterns (Ye et al., 2021). In addition, there are significant differences in development between the different countries and regions. In studies taking Chinese cities as experimental areas, the FoM values are mostly concentrated between 0.1 and 0.2 (Chen et al., 2016; Liang et al., 2018a). This is similar to the simulation accuracy observed in this study.

### Table 9. Interval division method of the similarity coefficient.

| Interval level | Similarity coefficient interval |
|----------------|---------------------------------|
| Higher similarity | $\mu + 0.5\text{std}$                           |
| High similarity | $\mu + \text{std}$                           |
| Medium similarity | $\mu$                                        |
| Low similarity  | $\mu - \text{std}$                           |
| Lower similarity | $\mu - 0.5\text{std}$                      |

### Figure 9. FoM and the number of geographic cases

Simulation accuracy after the constraint improved by 1.6%, confirming the effectiveness of the constraint method.

The main principle of the quantity demand constraint is that the smaller the similarity coefficient, the stronger the correlation between cases. Therefore, the effect of the constraint can be analyzed by comparing the accuracy of different similarity coefficient intervals. The mean-standard deviation method is used to divide the similarity coefficient interval of the urbanized simulated cases into five grades (Table 9).

We counted the correct and incorrect cases in each interval. As shown in Figure 10(a), the high and medium similarity levels had the most correct simulated cases, followed by higher similarity level. However, the incorrect cases in the three grades also showed an increasing state as the grade dropped. In addition, the fewest correct simulations were in the low and lower similarity levels, with the most incorrect cases appearing in the lower level. Simulation accuracy and the FoM index can show the relation between interval grade and accuracy more directly. According to Figure 10(b), urban accuracy, overall accuracy, and the FoM index all showed an obvious downward trend with a decrease in interval grade. The similarity interval grade had an obvious positive correlation with the simulation accuracy. It also showed that the smaller the similarity coefficient, the higher the similarity between cases and the higher the reliability of the simulation results. Therefore, the quantity demand method was proven to be effective.

### Figure 10. The relationship between simulation accuracy and coefficient interval

TheFoM is above 0.40. According to the distribution and shape of urban land, most of Tehran’s urban land is concentrated in the north, with an overall high degree of urban concentration, and the southern part of the city is the main direction of urban growth; Mumbai’s urban-land connectivity is relatively high, and urban growth is mainly based on internal filling. However, the urban land in Jixi city is mainly scattered in the south, middle, and east of the study area, with more directions and more complicated urban growth patterns (Ye et al., 2021). In addition, there are significant differences in development between the different countries and regions. In studies taking Chinese cities as experimental areas, the FoM values are mostly concentrated between 0.1 and 0.2 (Chen et al., 2016; Liang et al., 2018a). This is similar to the simulation accuracy observed in this study.
Simulation accuracy and efficiency

According to the experimental results (Figure 9), with the increase in the number of sample cases, the probability of retrieving more similar cases can be improved, and the reliability of the simulation results can be enhanced. However, this also means that the accuracy is limited by the number of historical cases. Specifically, the simulation accuracy is not satisfactory when the number of historical cases is small. At the same time, the number of cases also affects the operation speed of the model. If excessive historical cases are collected to improve the accuracy, this usually leads to longer simulation times, which benefits neither the analysis nor the exploration of the simulation process. Therefore, it is difficult to balance the model’s simulation accuracy and efficiency under the influence of the number of historical cases, making a standard for sampling historical case numbers necessary. Future research can explore the influence of different historical periods on the simulation results to develop a method of increasing historical experience through multi-period data sampling and to explore the correlation between simulation accuracy and efficiency to perfect the model’s reasoning in urban-growth research.

Conclusions

CBR simplifies knowledge acquisition and relies on sufficient historical data to realize quantitative analyses and phenomena predictions. It can effectively solve the barriers of conversion rules that emerge in the process of spatio-temporal land-use change simulation. This study presents a UGSCBR model for urban-growth simulation. By introducing a three-step process, this model improved the shortcomings of the case structures and model algorithms traditionally used in research (Du et al., 2010). It also adapts CBR reasoning to the urban-growth simulation process and solves the problem of CBR not being sensitive to time scales. This study applied the model to an urban-growth simulation of Jixi city from 2010 to 2015, and the experiment showed the flexibility and simplicity of the UGSCBR model. This model provides a new direction for the study of urban growth.

Acknowledgments

We thank the reviewers for their insightful and intelligent suggestions on the manuscript.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the University Nursing Program for Young Scholars with Creative Talents in Heilongjiang Province under Grant number UNPYSCST-2020032, and the Natural Science Foundation of Heilongjiang Province of China under Grant number YQ2019D006.

ORCID

X. Ye (http://orcid.org/0000-0002-1432-2786)

Data and code availability statement

The data and codes that support the findings of this study are available with identifiers at a private link (https://figshare.com/s/69e1316ae70ac4bace4e).

References

Aamodt, A., & Plaza, E. (1994). Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI Communications*, 7(1), 39–59. https://doi.org/10.3233/AIC-1994-7104

Aburas, M. M., Ho, Y. M., Ramli, M. F., & Ash’aari, Z. H. (2017). Improving the capability of an integrated CA-Markov model to simulate spatio-temporal urban growth trends using an analytical hierarchy process and frequency ratio. *International Journal of Applied Earth Observation and Geoinformation*, 59, 65–78. https://doi.org/10.1016/j.jag.2017.03.006

Ahmad, F., Goparaju, L., & Qayum, A. (2017). ULC analysis of urban spaces using Markov chain predictive model at Ranchi in India. *Spatial Information Research*, 25(3), 1–9. https://doi.org/10.1007/s41324-017-0102-x

Blanchard, P., Dawin, J. R., & Volchenkov, D. (2010). Markov chains or the game of structure and chance. *The European Physical Journal. Special Topics*, 184(1), 1–82. https://doi.org/10.1140/epjst/e2010-01232-1

Boughorbel, S., Jarray, F., & El-Anbari, M. (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLoS One*, 12(6), e0177678. https://doi.org/10.1371/journal.pone.0177678

Campos, P. B. R., de Almeida, C. M., & de Queiroz, A. P. (2018). Educational infrastructure and its impact on urban land use change in a peri-urban area: A cellular-automata based approach. *Land Use Policy*, 79, 774–788. https://doi.org/10.1016/j.landusepol.2018.08.036

Chen, Y., Li, X., Liu, Y., & Ai, B. (2013). Modeling urban land-use dynamics in a fast developing city using the modified logistic cellular automaton with a patch-based simulation strategy. *International Journal of Geographical Information Science*, 28(2), 234–255. https://doi.org/10.1080/13658816.2013.831868

Chen, Y., Li, X., Liu, X., Ai, B., & Li, S. (2016). Capturing the varying effects of driving forces over time for the simulation of urban growth by using survival analysis and cellular automata. *Landscape and Urban Planning*, 152, 59–71. https://doi.org/10.1016/j.landurbplan.2016.03.011

Cheng, J., & Masser, I. (2004). Understanding spatial and temporal processes of urban growth: Cellular automata modelling. *Environment and Planning. B. Planning & Design*, 31(2), 167–194. https://doi.org/10.1068/b2975
Coppedge, B. R., Engle, D. M., & Fuhlendorf, S. D. (2007). Markov models of land cover dynamics in a southern Great Plains grassland region. Landscape Ecology, 22(9), 1383–1393. https://doi.org/10.1007/s10980-007-9116-4

Dou, J., Chang, K. T., Chen, S., Ali, Y., Liu, J. K., Xia, H., & Zhu, Z. (2015). Automatic case-based reasoning approach for landslide detection: Integration of object-oriented image analysis and a genetic algorithm. Remote Sensing, 7(4), 4318–4342. https://doi.org/10.3390/rs70404318

Du, Y., Wen, W., Cao, F., & Ji, M. (2010). A case-based reasoning approach for land use change prediction. Expert Systems with Applications, 37(8), 5745–5750. https://doi.org/10.1016/j.eswa.2010.02.035

Du, Y., Wu, D., Liang, F., & Li, C. (2013). Integration of case-based reasoning and object-based image classification to classify SPOT images: A case study of aquaculture land use mapping in coastal areas of Guangdong province, China. GIScience & Remote Sensing, 50(5), 574–589. https://doi.org/10.1080/15481603.2013.842292

Du, Y., Yong, G., Lakhan, V. C., Sun, Y., & Feng, C. (2012). Comparison between CBR and CA methods for estimating land use change in Dongguan, China. Journal of Geographical Sciences, 22(4), 716–736. https://doi.org/10.1007/s11442-012-0958-6

Du, Y., Zhou, C., Shao, Q., Su, F., Shi, Z., & Ye, S. (2002). Theoretic and application research of geo-case-based reasoning. Acta Geographica Sinica, 57(2), 151–158. https://doi.org/10.3321/j.0377-5442.2002.02.004

Duncan, E. E., Kuma, J. S., & Primpong, S. (2009). Open pit mining and land use changes: An example from Bogosu-Prestea area, South West Ghana. The Electronic Journal of Information Systems in Developing Countries, 36(1), 1–10. https://doi.org/10.1002/j.1681-4835.2009.tb00250.x

Firozjahi, M. K., Sedighi, A., Argany, M., Jelokhani-Niaraki, M., & Arsanjani, J. J. (2019). A geographical direction-based approach for capturing the local variation of urban expansion in the application of CA-Markov model. Cities, 93, 120–135. https://doi.org/10.1016/j.cities.2019.05.001

Holt, A. (1999). Applying case-based reasoning techniques in GIS. International Journal of Geographical Information Science, 13(1), 9–25. https://doi.org/10.1080/136588199214436

Jafari, M., Majedi, H., Monavari, S. M., Alesheikh, A. A., & Zarkesh, M. K. (2017). Dynamic simulation of urban expansion through a CA-Markov model case study: Hyrcanian region, Gilan, Iran. European Journal of Remote Sensing, 49(1), 513–529. https://doi.org/10.5721/EuJRS20164927

Kantakumar, L. N., Kumar, S., & Schneider, K. (2019). SUSM: A scenario-based urban growth simulation model using remote sensing data. European Journal of Remote Sensing, 52(sup2), 26–41. https://doi.org/10.1080/22797254.2019.1585209

Li, X., & Liu, X. (2006). An extended cellular automation using case-based reasoning for simulating urban development in a large complex region. International Journal of Geographical Information Science, 20(10), 1109–1136. https://doi.org/10.1080/13658810600816876

Li, X., Liu, X., Li, X., Chen, Y., Tian, H., & Yao, Y. (2018b). Delineating multi-scenario urban growth boundaries with a CA-based FLUS model and morphological method. Landscape and Urban Planning, 177, 47–63. https://doi.org/10.1016/j.landurbplan.2018.04.016

Li, X., Liu, X., Li, D., Zhao, H., & Chen, G. (2018a). Urban growth simulation by incorporating planning policies into a CA-based future land-use simulation model. International Journal of Geographical Information Science, 32(11), 2294–2316. https://doi.org/10.1080/13658816.2018.1502441

Liao, Z., Zhou, C., Tian, W., Hu, T., & Guo, R. (2019). CBR-based integration of a hydrodynamic and water quality model and GIS—A case study of Chaohu City. Environmental Science and Pollution Research, 26(7), 6436–6449. https://doi.org/10.1007/s11356-018-3862-5

Liu, X., Lei, M., Xia, L., Ai, B., Li, S., & He, Z. (2013). Simulating urban growth by integrating landscape expansion index (LEI) and cellular automata. International Journal of Geographical Information Science, 28(1), 148–163. https://doi.org/10.1080/13658816.2013.831097

Liu, X., Liang, X., Li, X., Xu, X., Ou, J., Chen, Y., Li, S., Wang, S., & Pei, F. (2017). A future land use simulation model (FLUS) for simulating multiple land use scenarios by coupling human and natural effects. Landscape and Urban Planning, 168, 94–116. https://doi.org/10.1016/j.landurbplan.2017.09.019

Liu, W., Wang, S., Zhou, Y., Wang, L., Zhu, J., & Wang, F. (2015). Lightning-caused forest fire risk rating assessment based on case-based reasoning: A case study in DaXingAn Mountains of China. Natural Hazards, 81(1), 347–363. https://doi.org/10.1007/s11069-015-2083-1

Liu, Z., Yang, Y., He, C., & Tu, M. (2019). Climate change will constrain the rapid urban expansion in drylands: A scenario analysis with the zoned land use scenario dynamics-urban model. Science of the Total Environment, 651(Pt 2), 2772–2786. https://doi.org/10.1016/j.scitotenv.2018.10.177

Machado, D. F. T., de Menezes, M. D., Silva, S. H. G., & Curi, N. (2019). Transferability, accuracy, and uncertainty assessment of different knowledge-based approaches for soil types mapping. Catena, 182, 104134. https://doi.org/10.1016/j.catena.2019.104134

Mantaras, R. L. D., Mcherry, D., Bridge, D., Leake, D., Smyth, B., Craw, S., Faltings, B., Maher, M. L., Cox, M. T., Forbus, K., Keane, M., Aamodt, A., & Watson, L. (2006). Retrieval, reuse, revision and retention in case-based reasoning. The Knowledge Engineering Review, 20(3), 215–240. https://doi.org/10.1017/S0269888906000646

Mookherjee, D., & Hoerauf, E. (2015). Is it sprawling yet? A density-based exploration of sprawl in the urban agglomeration region around the mega city of Delhi. Urban Development Challenges, Risks and Resilience in Asian Mega Cities, 49–60. https://doi.org/10.978-4-431-55042-6

Ouyang, X., Wang, Z., & Zhu, X. (2019). Construction of the ecological security pattern of urban agglomeration under the framework of supply and demand of ecosystem services using Bayesian network machine learning: Case study of the Changsha–Zhuzhou–Xiangtan urban agglomeration, China. Sustainability, 11(22), 6416. https://doi.org/10.3390/su11226416

Pontius, R. G., Jr, Hufnaker, D., & Denman, K. (2004). Useful techniques of validation for spatially explicit land-change models. Ecological Modelling, 179(4), 445–461. https://doi.org/10.1016/j.ecolmodel.2004.05.010

Qian, J. P., Li, X., Bin, A., Yeh, A. G., Chen, X. Y., Liu, K., & Zhuo, L. (2007). Land-use short-term change detection based on time-series-case-based-reasoning. Journal of Natural Resources, 22(5), 735–746. https://doi.org/10.3321/j.1000-3037.2007.05.008
Schueler, V., Kuemmerle, T., & Schroder, H. (2011). Impacts of surface gold mining on land use systems in western Ghana. *Ambio, 40*(5), 528–539. https://doi.org/10.1007/s13280-011-0141-9

Shafizadeh-Moghadam, H., Asghari, A., Tayyebi, A., & Taleai, M. (2017). Coupling machine learning, tree-based and statistical models with cellular automata to simulate urban growth. *Computers, Environment and Urban Systems, 64*, 297–308. https://doi.org/10.1016/j.compenvurbsys.2017.04.002

Shafizadeh-Moghadam, H., Hagenauer, J., Farajzadeh, M., & Helbich, M. (2015). Performance analysis of radial basis function networks and multi-layer perceptron networks in modeling urban change: A case study. *International Journal of Geographical Information Science, 29*(4), 606–623. https://doi.org/10.1080/13658816.2014.993989

Shafizadeh-Moghadam, H., Minaei, M., Pontius, R. G., Jr, Asghari, A., & Dadashpoor, H. (2021). Integrating a forward feature selection algorithm, random forest, and cellular automata to extrapolate urban growth in the Tehran-Karaj region of Iran. *Computers, Environment and Urban Systems, 87*, 101595. https://doi.org/10.1016/j.compenvurbsys.2021.101595

Tan, Y., Jiao, L., Shuai, C., & Shen, L. (2018). A system dynamics model for simulating urban sustainability performance: A China case study. *Journal of Cleaner Production, 199*, 1107–1115. https://doi.org/10.1016/j.jclepro.2018.07.154

Watson, I. (1999). Case-based reasoning is a methodology not a technology. *Knowledge-Based Systems, 12*(5–6), 303–308. https://doi.org/10.1016/S0950-7051(99)00020-9

Wu, D., & Ning, S. (2018). Dynamic assessment of urban economy-environment-energy system using system dynamics model: A case study in Beijing. *Environmental Research, 164*(JUL), 70. https://doi.org/10.1016/j.envres.2018.01.029

Ye, X., Dong, L., Lv, L., & Shang, Y. (2021). Spatiotemporal evolution law and driving force of mining city patterns. *Environmental Science and Pollution Research, 29*(7), 10291–10307. https://doi.org/10.1007/s11356-021-16488-5

Yu, F., & Li, X. Y. (2018). Improving emergency response to cascading disasters: Applying case-based reasoning towards urban critical infrastructure. *International Journal of Disaster Risk Reduction, 30*, 244–256. https://doi.org/10.1016/j.ijdrr.2018.04.012

Yu, F., Li, X. Y., & Han, X. S. (2018). Risk response for urban water supply network using case-based reasoning during a natural disaster. *Safety Science, 106*, 121–139. https://doi.org/10.1016/j.ssci.2018.03.003

Zhang, J., Rao, Y., Geng, Y., Fu, M., & Prischepov, A. V. (2017). A novel understanding of land use characteristics caused by mining activities: A case study of Wu’an, China. *Ecological Engineering, 99*, 54–69. https://doi.org/10.1016/j.ecoleng.2016.11.026