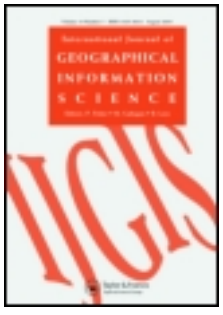

International Journal of Geographical Information Science

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tgis20

Calibrating a cellular automata model for understanding rural-urban land conversion: a Pareto front-based multi-objective optimization approach

Kai Cao[ab], Bo Huang[c], Manchun Li[d] & Wenwen Li[e]

[a] Center for Geographic Analysis, Harvard University, Cambridge, MA, USA

[b] World History Center, University of Pittsburgh, Pittsburgh, PA, USA

[c] Department of Geography and Resource Management, The Chinese University of Hong Kong, Shatin, NT, HongKong

[d] Department of Geographic Information Science, Nanjing University, Nanjing, PR China

[e] GeoDaCenter for Geospatial Analysis and Computation, School of Geographical Sciences and Urban Planning, Arizona State University, Tempe, AZ, USA

Published online: 05 Dec 2013.

To cite this article: Kai Cao, Bo Huang, Manchun Li & Wenwen Li , International Journal of Geographical Information Science (2013): Calibrating a cellular automata model for understanding rural-urban land conversion: a Pareto front-based multi-objective optimization approach, International Journal of Geographical Information Science, DOI: 10.1080/13658816.2013.851793

To link to this article: http://dx.doi.org/10.1080/13658816.2013.851793

International Journal of Geographical Information Science, 2013
http://dx.doi.org/10.1080/13658816.2013.851793

Taylor & Francis
Taylor & Francis Group

Calibrating a cellular automata model for understanding rural–urban land conversion: a Pareto front-based multi-objective optimization approach

Kai Cao[a,b]*, Bo Huang[c], Manchun Li[d] and Wenwen Li[e]

[a]*Center for Geographic Analysis, Harvard University, Cambridge, MA, USA;* [b]*World History Center, University of Pittsburgh, Pittsburgh, PA, USA;* [c]*Department of Geography and Resource Management, The Chinese University of Hong Kong, Shatin, NT, HongKong;* [d]*Department of Geographic Information Science, Nanjing University, Nanjing, PR China;* [e]*GeoDaCenter for Geospatial Analysis and Computation, School of Geographical Sciences and Urban Planning, Arizona State University, Tempe, AZ, USA*

(*Received 18 June 2013; final version received 1 October 2013*)

Cellular automata (CA) modeling is useful to assist in understanding rural–urban land conversion processes. Although CA calibration is essential to ensuring an accurate modeling outcome, it remains a significant challenge. This study aims to address that challenge by developing and evaluating a multi-objective optimization model that considers the objectives of minimizing minus maximum likelihood estimation (MLE) value and minimizing number of errors (NOE) when calibrating CA transition rules. A Pareto front-based heuristic search algorithm, the Non-dominated Sorting Genetic Algorithm-II (NSGA-II), is used to obtain optimal or near-optimal solutions. The proposed calibration approach is validated using a case study from New Castle County, Delaware, United States. A comparison of the NSGA-II-based calibration model, the generic Logit regression calibration approach (MLE-based Generic Genetic Algorithm (GGA) calibration approach), and the NOE-based GGA calibration approach demonstrates that the proposed calibration model can produce stable solutions with better simulation accuracy. Furthermore, it can generate a set of solutions with different preferences regarding the two objectives which can provide CA simulation with robust parameters options.

Keywords: NSGA-II; land conversion; rural–urban; cellular automata; calibration; Logit regression

1. Introduction

As a dynamic, bottom-up modeling technique, cellular automata (CA) are suitable for modeling complex spatio-temporal dynamics such as those used in studies of land-use change and urban growth (White and Engelen 1993, 1997, Batty *et al*. 1994a, Clarke *et al*. 1997, Clarke and Gaydos 1998, Wu 1998, Wu and Webster 1998, Batty *et al*. 1999, Li and Yeh 2000, Silva and Clarke 2002, Pontius and Malanson 2005, Li and Liu 2007, Li *et al*. 2007, Yang 2008, Feng and Liu 2012, 2013). However, the specification of CA state transition rules remains a significant research challenge, despite the emergence of CA as a powerful modeling tool in urban growth simulation (Batty 1998). The calibration of CA for urban growth modeling requires finding the best combination of transition rule values for matching the modeled urban phenomenon to its real counterpart. Calibration is critical

*Corresponding author. Email: kaicao@pitt.edu

to the performance of a CA model (Batty *et al*. 1994a, 1994b, Landis and Zhang 1998, Batty *et al*. 1999).

The difficulty in calibrating CA rules is mainly due to the large, unknown solution search space and its exponential growth as more complicated and larger number of variables become involved. A number of CA calibration methods have been developed for urban growth modeling and have achieved various levels of success and efficiency. Clarke *et al*. (1997, 1998) calibrated the CA model by using one visual and four statistical tests to find the best parameter values. Wu and Webster (1998) defined the CA transition rules according to a multi-criteria evaluation (MCE) formulation. To remove the subjective parameter setting, some studies based on Logit regression models have also been conducted. Wu (2002), for instance, successfully calibrated a CA model by combining global and local development probability with the regression model and considering neighborhood rules. Although the statistics-based methods, such as the Logit regression model, can identify better parameters for CA modeling, their drawback is that regression models are based on the assumption of independent variables and samples, which are often hard to match with real-world cases. From another perspective, CA models must be assessed according to plausibility (Batty 1998) rather than these statistical tests, which means that the target of the rural–urban CA model is to capture the basic features regarding the precision of rural–urban land-use change. Thus, a gap exists between the statistical parameters estimation methods and the real-world precision of the calibrated rural–urban CA model.

In the last decade, some studies have applied methods developed in the artificial intelligence literature to calibrate the CA models, which directly aimed to pursue the precision of land-use change. Li and Yeh (2002) calibrated a CA model using a neural network algorithm. A training set and its corresponding modeling output were used to train the neural network to reproduce future urban land-use patterns. Genetic algorithms (GAs) have also been used to successfully calibrate CA models (Colonna *et al*. 1998, Goldstein 2003, Yang and Li 2007, Yang *et al*. 2007, Shan *et al*. 2008, Feng and Liu 2012). GAs are based on biological principles to direct the search toward regions of solution space that contain likely improvements (Goldberg 1989). Hence, the application of GAs might offer better precision in land conversion simulation. The state-of-the-art design of GA fitness functions for calibrating CA models is often informed by maximizing the percentage of correctness (goodness of fit) or minimizing the number of errors (NOE), which may introduce drawbacks including existing unreasonably calibrated parameters and weak robustness. There may also be a need for improvement in the use of GAs to calibrate CA models, not only in relation to the design of fitness function, but also in relation to the core design of GA form. Herein, the Non-dominated Sorting Genetic Algorithm-II (NSGA-II), as one type of Pareto front-based GA model, which has also been utilized to solve the spatial land-use optimization problem (Cao *et al*. 2011), is utilized in accordance with two individual objectives – the minimization of the minus maximum likelihood estimation (MLE) value (statistical meaning) and the minimization of NOE – to effectively, efficiently, and stably calibrate the parameters of CA models.

The remainder of this paper begins with a brief introduction to CA and the rural–urban conversion model used. Then, NSGA-II is explained, along with its desirable ability to calibrate CA transition rules through the integration of a fitness function based on an MLE (Logit regression model) as one objective and NOE for reflecting simulation results as the other one. Compared to a generic Logit regression calibration approach (an MLE-based generic genetic algorithm (GGA) calibration approach) (GLRCA) and the NOE-based GGA calibration approach, the NSGA-II-based CA calibration model is validated by a

case study from New Castle County, Delaware, United States, in which the rural–urban land-use change from 1992 to 1997 was simulated.

2. CA for rural–urban conversion modeling

CA was originally introduced by Ulam and von Neumann in the 1940s to study the behavior of complex systems (VonNeumann and Burks 1966). It is a dynamic discrete model studied in computability theory, mathematics, physics, complexity science, theoretical biology, and microstructure modeling. It consists of cells or pixels, the states (such as land-use types), neighborhoods, and transition rules. The future states of the cells are changed from the current states to the specific states under the consideration of neighborhood influences and transition rules. Undoubtedly, the setting of the transition rules is one of the essential issues in CA modeling.

CA has been successfully applied to simulations of the rural–urban conversion process in previous decades. The generic CA for modeling rural–urban conversion could be formulated as follows:

$$S_{t+1}(i) = \begin{cases} \text{Change to Urban Cell, } \textit{If rules satisfied} \\ \text{Keep the Status of Rural Cell, } \textit{Otherwise} \end{cases} \tag{1}$$

where $S_{t+1}(i)$ is the state of the cell i at time $t+1$.

The general transition rules can be divided into the following parts:

$$P_i^{t+1} = f(pN_i^t, pF_i^t, pC_i^t, pO_i^t) \tag{2}$$

where P_i^{t+1} is the transition preference value of the transition rules, pN_i^t is the transition preference value of neighborhood influence, pF_i^t is the transition preference value of global factors, pC_i^t is the transition preference value of constraints, and pO_i^t is the transition preference value of random influence of other driving forces. The function of Equation (2) is flexible about how to combine the above-mentioned probabilities. The sum of these probabilities is applied in this study.

The transition preference value of neighborhood influence reflects the interactions between cells and their neighborhoods, along with the characteristics of bottom-up self-organization evolution. With regard to the 3 × 3 cells kernel, the neighborhood potentiality of cell transition is defined as follows:

$$pN_i^t = \frac{\sum_{3\times 3} con(S_t(i) = urban)}{3 \times 3 - 1} \tag{3}$$

where $S_t(i)$ is the state of the cell i at time t, $con()$ is a conditional function that returns 1 if $S_t(i)$ is urban land use. In this simulation, the neighborhood has been defined as eight immediately neighboring cells, as in previous studies (Fulong Wu 2002, Yang 2008). It must be noted that pN_i^t is denominated by time t, which means that it will change along with the simulation.

The transition preference value of global factors reflects the effects of natural and socioeconomic conditions on urban development. It generally includes factors such as population density, distance to roads, distance to developed center, distance to railway,

slope, and ecological suitability. Mathematically, this can be generalized as the estimation of probability of particular state transition occurring at a particular location i through a function of global factors $(x_1, x_{2,}..., x_n)$. A Logit model can be developed to calculate the probability of development as follows:

$$pF_i^t = \frac{1}{1 + \exp(-(a_0 + a_1x_1 + a_2x_2 + ... + a_nx_n))} \tag{4}$$

where $(x_1, x_{2,}..., x_n)$ are the global factors (variables), $(a_1, a_{2,}..., a_n)$ are the coefficients of the Logit regression model, and a_0 is a constant.

Constraint transition preference value is used to represent the influence of the natural constraints and the high-level land-use planning on urban development. The formula can be specified as follows:

$$pC_i^t = con(s_t(i) \neq restrict) \tag{5}$$

where $S_t(i)$ is the state of the cell i at time t, $con()$ is a conditional function that returns 1 if $S_t(i)$ is not restricted.

Due to the uncertainty of urban development, it is necessary to bring forward a random parameter to control the simulation process (Li and Yeh 1999).

$$pO_i^t = 1 + (-\ln\gamma)^\alpha \tag{6}$$

where γ is a random value within the range of 0 and 1, α is an integer within the range of 1 and 10 for controlling the influence of other uncertain factors.

All the four aspects mentioned above are very essential for the entire model. In this research, we are focusing on the calibration of pF_i^t, since it involves the most factors that need to be calibrated, with an assumption that all other aspects use the same parameters setting. Thus, this study highlights the potential improvement of the calibration of pF_i^t to construct more accurate, suitable, and stable CA-based rural–urban conversion models.

3. NSGA-II algorithm for calibration

With regard to the limitations of the early calibration methods and the characteristics of CA-based rural–urban conversion models, the NSGA-II multi-objective optimization scheme-based calibration method, as one kind of Pareto front-based optimization models, can offer improvement by integrating the fitness function of the generic Logit regression model and the percentage of correction for the simulation results. In this section, the objectives are introduced at first, followed by the principles of NSGA-II, which are related to its effective searching capability of Pareto front solutions (calibrated parameters sets). And the details of NSGA-II for CA rural–urban conversion model calibration are also explained as well as the calibration of development probability and validation method.

3.1. Objectives

3.1.1. MLE

With respect to the generic Logit regression model, the MLE method is typically used to estimate the parameters during the regression process.

The regression model is as mentioned in Equation (4). Accordingly, the likelihood function can be defined as follows:

$$L(\beta) = \prod_{i=1}^{n} pF_i^t y_i [1 - pF_i^t]^{1-y_i} \tag{7}$$

The logarithm likelihood value should be as follows:

$$LL(\beta) = \ln[L(\beta)] = \sum_{i=1}^{n} \left\{ y_i \ln[pF_i^t] + (1 - y_i) \ln[1 - pF_i^t] \right\} \tag{8}$$

where n is the number of samples, and $(y_1, y_2, \ldots, y_n)$ denote observations.

Because minimization operation is utilized in this optimization process, minus $LL(\beta)$ will be the final fitness function for reflecting the objective value of MLE.

3.1.2. NOE

With regard to NOE, which could reflect the precision of the simulation results, the objective is as follows.

In Table 1, A is the number of cells under a situation in which the observation value is the same as the simulated value and the cells did not change. B is the number of cells under a situation in which the simulated value changes, but the real value does not. C is the number of cells under a situation in which the simulated value does not change, but the real value does. D is the number of cells under a situation in which the observation value is the same as the simulated value and the cells changed. P1 is equal to $A/(A + B)$; P2 is equal to $D/(C + D)$; P3 is equal to $A/(A + C)$; P4 is equal to $D/(B + D)$; and P is equal to $(A + D)/(A + B + C + D)$.

Obviously, NOE could be calculated by $(B + C)$.

3.2. *Principles of NSGA-II*

NSGA-II, developed by Deb *et al*. (2002) as an improved version of NSGA, is an efficient multi-objective evolutionary algorithm that uses an elitist approach that sorts the population at different 'fronts' using a non-dominated sorting method with a particular bookkeeping strategy. The crowding distance sorting is another essential part in ranking the population, at which point the best individuals in terms of non-dominance and diversity are chosen. A sketch of the algorithm that indicates how a solution P_t is progressed to P_{t+1} through the front using the crowding distance sorting is shown in Figure 1.

Table 1. Conversion matrix.

		Simulated		
		Non change	Change	%
Real	Non change	A	B	P1
	Change	C	D	P2
%		P3	P4	P

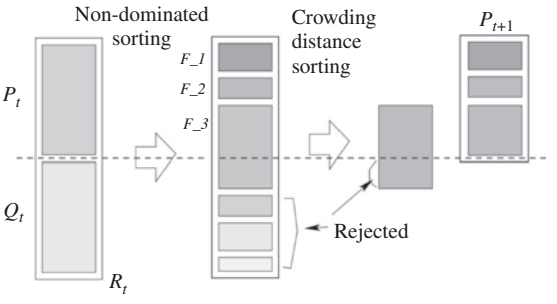

Figure 1. A sketch of NSGA-II (Deb et al. 2002; Cao et al 2011).

3.2.1. *Non-dominated sorting*

To sort a population of size N according to the level of non-domination, each solution must be compared to every other solution in the population to check whether it is dominated. This requires an $o(ON)$ computation, where ON in the brackets stands for the number of objectives. For enumeration to reflect the entire first Pareto front, $o(ON^2)$ comparisons are required, whereas in the worst situations, the computation to obtain all of the fronts, level by level, requires $o(ON^3)$ comparisons. Within NSGA-II, the book-keeping strategy can be used to decrease the required computations to $o(ON^2)$ at most.

3.2.2. *Crowding distance*

The crowding distance is another essential concept proposed by Deb *et al.* (2000) for the NSGA-II algorithm, the target of which is to generate an estimation of the density of solutions surrounding a particular solution within the population. The crowding distance for a point i is the estimate of the size of the largest cuboid enclosing the point i, but including no other point in the population. It calculates the average distance between two points on either side of this point along the objective axes (as shown in Figure 2).

The details could be seen in the paper of Deb *et al.* (2002).

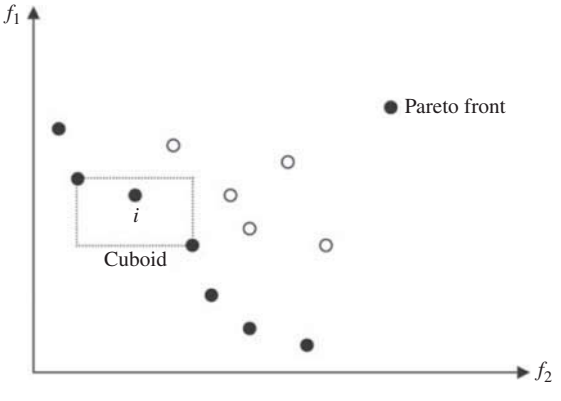

Figure 2. The crowding distance calculation (Deb et al. 2002; Cao et al. 2011).

3.3. *Details of NSGA-II model for CA rural–urban conversion model calibration*

NSGA-II is an algorithm with a particularly good Pareto front search capability that generates sufficiently diverse solutions (Deb *et al*. 2002). Figure 3 illustrates the general process of the parameters calibration for defining the transition rules of a CA model by NSGA-II.

3.3.1. *The representation and initialization of population*

GAs typically cannot directly handle optimization without representation (encoding), which is the first step in the application. Each genotype in the population represents a complete specification for a single solution, and the genotypes are made up of genes that should be specified by the individual components of a solution. The earliest popular representation for GA implementation was the fixed-length binary string. This remains the most flexible and popular representation so far, since it is straightforward to explain and provide further crossover and mutation operations. There are other additional representation methods for specific cases, such as real value, percentage and priority (Matthews 2001), grid, and quad-tree. In this study, the simple parameters are easily represented by the fixed-length binary string. The 'gray code' is also utilized due to its characteristic that any two neighboring codes only have one different character, which improves the reasonability of the crossover and mutation process.

The initialization, which influences both the convergence precision and efficiency, is another essential step of GA. To keep the range of the initialized solutions and the precision of the iteration in this study, the randomly created populations were the initialized population.

In this study, each gene takes 10 bits. And referring to another case of applying NSGA-II on the field of geographic optimization problem (Cao *et al*. 2011) as well as a few experiments, the total size of the population is set to be 100.

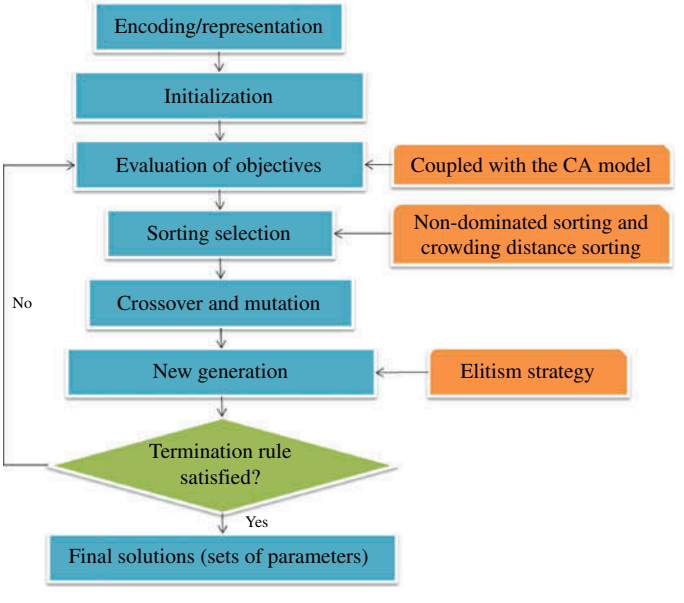

Figure 3. Flowchart of NSGA-II for calibrating a CA Model.

3.3.2. Sorting selection and elitism

Selection, also named as reproduction, is the process of selecting the better individuals based on the fitness function. The reproduction operator is used to choose the survival of the fittest individuals; that is, fitness is the criterion, and individuals with high fitness will have a higher probability of surviving into the next generation. On the contrary, those with low fitness will have a lower probability of surviving. In this research, the selection operation is based on the above-mentioned principles, non-dominated sorting and crowding distance, which embody the essential parts of NSGA-II. Elitism is another parameter capable of improving the iteration process. In elitism selection, the top 10% of the strings would be duplicated directly to the next generation. This step is performed to retain good solutions suitable to the current generation, which avoids missing the best solutions found in early iterations.

3.3.3. Crossover and mutation

Nature produces the next generation using a mating process. This is accomplished by two parents creating offspring that contain genetic material from both parents. Crossover (also named as recombination) is when two individuals with a certain probability are chosen and exchange one or some of their parts. The offspring generated by this process retain the basic characteristics of the individual parents. The key issues in this process are deciding the point of crossover and performing the exchange between the parents. This is the essential characteristic of GAs, which differs from other forms of evolutionary computation.

Matching is the inevitable prerequisite for the crossover process. The popular matching method is random while the real crossover process involves swapping the matching pairs. The general crossover operators include one-point, two-point, uniform, and arithmetic crossovers. The one-point crossover, also called a simple crossover, is utilized in this research and is illustrated in the following figure (Figure 4).

Similar to biological mutation, mutation in GAs maintains genetic diversity from one generation of a population to the next. It uses a small probability value to mutate some part or parts of an individual, such as swapping 1 and 0 in a binary-coded chromosome. Mutation itself is a kind of probability algorithm; however, when integrated with selection and crossover operation, the loss of useful information can be avoided. In this study, a simple mutation operator, which is similar to a simple crossover operator, is utilized to enhance the local searching capability of GAs, maintain the diversity of the individuals, and avoid premature convergence.

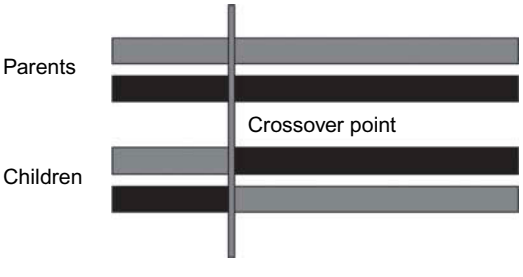

Figure 4. Single-point crossover (only one crossover point, randomly chosen, at which to perform the exchange of the chromosome pair).

3.4. Calibration of development probability and simulation

The calibration of development probability is carried out through repeating the above-mentioned genetic operations. The parameters of Equation (4) are targeted for calibration. First, the statistical analysis is used to obtain the significance of these factors and decide which factors should be considered for modeling and calibration. Second, the populations, with a size of 100, are randomly created for the following iteration. After that, the population is translated into the values used in the evaluation of the two objectives, MLE and NOE, operated by the NSGA-II model with its selection, crossover, and mutation. Finally, through 5000 iterations, which reach convergence, the calibrated parameters are obtained. Through integrating the four factors together and computing the comprehensive transition preference value, the simulation will be automatically finished by transiting the top certain number of rural land cells with higher comprehensive value to urban land cells.

3.5. Validation method for the calibration results

Besides parameters calibration, validation also plays an important role in modeling a complex system (Batty and Torrens 2005), which is also one necessary step to prove the meaning of the parameters calibrated during the calibration process. Many scholars discussed the validation topic and made great progress to bring us better and better validation models for land-use models (Kok *et al*. 2001, Brown *et al*. 2005, Vliet *et al*. 2011). But under the consideration of our emphasis, which mainly focuses on the calibration thought rather than a new land-use model, we utilize the following validation method. First, the calibration parameters obtained by the proposed algorithm are compared separately to the parameters obtained from GLRCA and NOE-based GGA calibration approach, which is only considering the accuracy of the simulation. Through these comparisons, the meaning of the two objectives considered in our proposed model could be clarified, and also the advantages of the proposed model could be demonstrated. On the other hand, the comparison of calibration results and their simulation precision based on a whole data set from the Pareto front and generic Logit regression solutions could finally prove the advance of the proposed model in this real case study.

The following section introduces a case study of New Castle County, Delaware, United States, in which the rural–urban states from 1992 to 1997 are simulated using the above-mentioned operations, thus validating the effectiveness of the proposed model.

4. Case study and validation of calibration results

4.1. Study area and data collection

New Castle County, covering about 1100 km^2, is one of three counties constituting the state of Delaware (Figure 5). It is the urban and manufacturing center of Delaware, and accommodates around 60% of the state's population. The urban land use is mainly located in the northern part of the county.

The data used in this study mainly include land use, terrain, and transportation network data. The land-use data were generated from digital orthophotos provided by the Delaware Office of State Planning Coordination (Figure 6). All land-use data were rasterized at a resolution of 50 × 50 m, which has totally 379,149 land cells. The original land uses were classified into five types: residential, commercial, industrial, agricultural, and others. The first three were categorized as urban areas. The agricultural lands were

Figure 5. Location of New Castle County in Delaware.

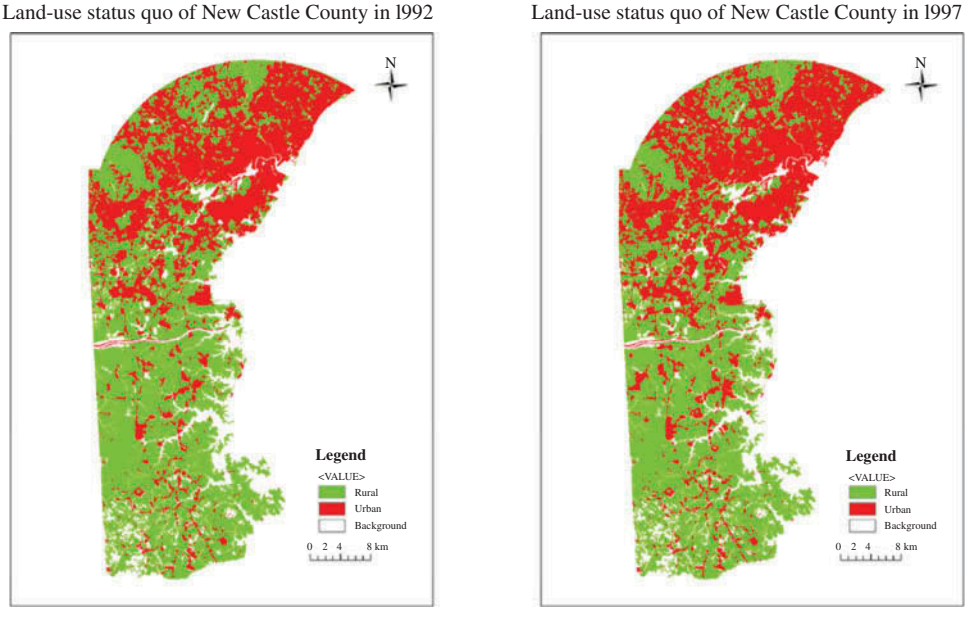

Figure 6. Land-use status quo of New Castle County in 1992 and 1997 (Huang et al. 2009).

defined as rural areas offering the potential for urban expansion, which is the essential feature in this study. The others were extracted from the research area since they include forest, water, and barren, which could be classified as unsuitable for the purpose of urban development.

The transportation network data were divided into two kinds: primary and small roads, each of which plays a different role in the urbanization process. The data were based on the Shapefile of the 2001 road network. As construction of the US transportation network was finalized during the 1960s, differences in it between 1992 and 2001 should be negligible.

The real land-use status quo of 1992 and 1997 can be seen in Figure 6; there are totally 14,999 land cells transited from rural to urban land use from 1992 to 1997, which would also be the target change numbers for the following validation of the case study in Section 4.4.

4.2. *Variables and sampling*

The rural cells of the land-use status quo in 1992 are the features studied here. As previously mentioned, pF_i^t is the essential formula used to show the independent variables and dependent variable. The dependent variable should be a binary variable that shows whether or not rural–urban conversion changed from 1992 to 1997, which can be determined by comparing the two land-use maps in Figure 6. Given the data limitations and the statistical significance test, the independent variables are as follows: (a) slope, (b) zoning suitability, (c) distance to industrial centers, (d) distance to small roads, and (e) distance to primary roads. Slope reflects the physical characteristic of land cells, which should influence rural–urban conversion. Zoning suitability indicates areas suitable for conversion to urban use. The Euclidean distances to the nearest industrial centers, small roads, and primary roads are calculated as the other three proximity variables affecting the rural–urban conversion. These proximity variables, derived using the ESRI ArcGIS software package, are shown in Figure 7 along with the first two variables.

In addition to pF_i^t, other factors also influence the comprehensive transition preference value during the simulation process. Because this study focuses on calibrating better parameters for computing the comprehensive transition preference value, and given that there are no other parameters in the other factors (pN_i^t, pC_i^t, and pO_i^t), we use the same pN_i^t, assuming that pC_i^t is as in Figure 8, and use 1 as pO_i^t to operate the rural–urban conversion simulation.

The cells that did not sustain land-use changes significantly outnumber the cells that did. To maintain the calibration accuracy, small, but important, areas must be represented in the samples (Congalton 1988). Therefore, a ratio between changed and unchanged rural cells should be specified. A ratio of two-thirds of the 5000 empirically and randomly chosen unchanged/changed rural cells could assure the calibration accuracy in this case.

4.3. *Parameter calibration based on NSGA-II and comparison with GLRCA and NOE-based GGA calibration approaches*

During the calibration process, the 5000 samples are used by the MLE-based GGA calibration approach, the NOE-based GGA calibration approach, and the NSGA-II-based calibration model. The calibration results are shown in Figures 9–11.

First, the GLRCA is used, and the convergence curves are as follows for the objectives of NOE and MLE by 500 iterations. The right curve in Figure 9 clearly shows the

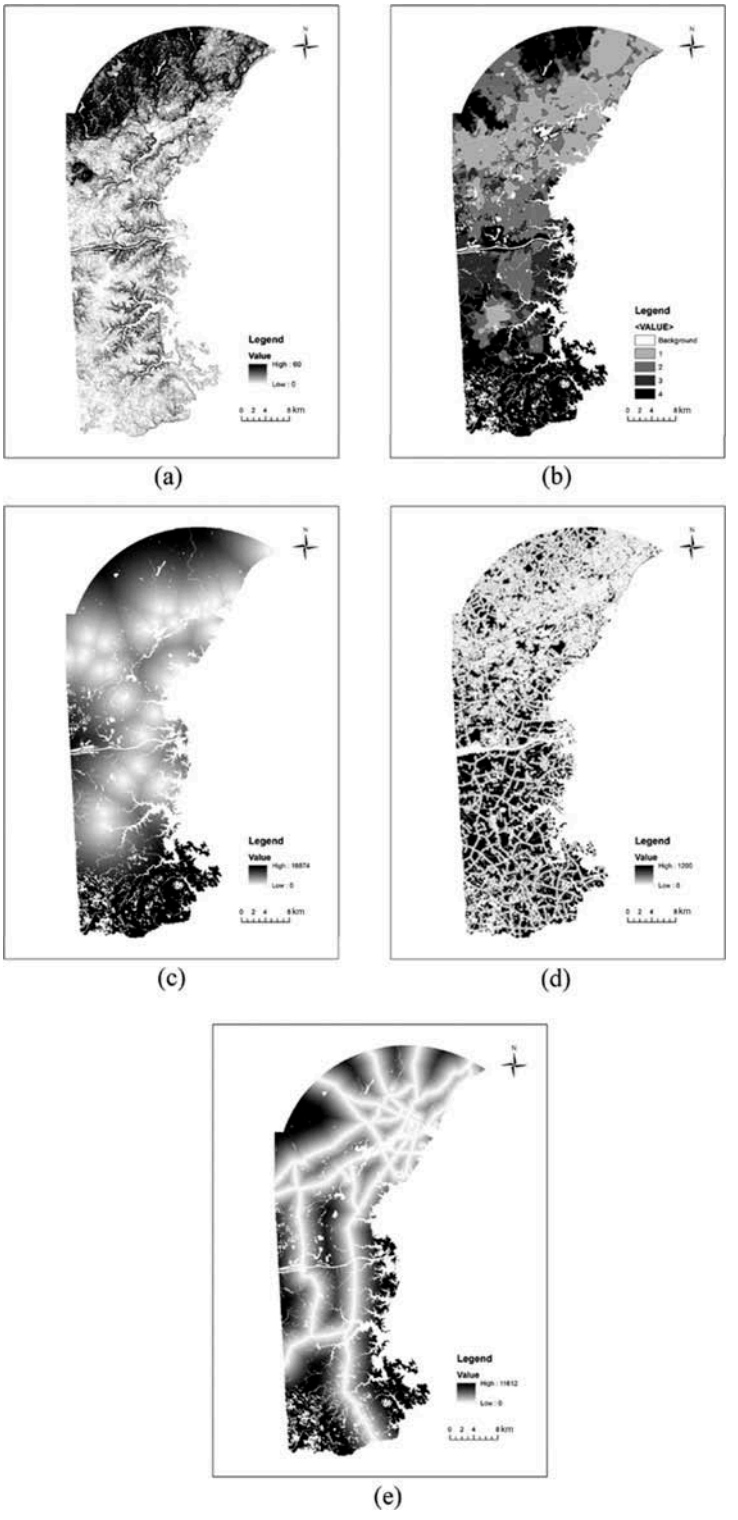

Figure 7. Predicted factors: (a) slope, (b) zoning suitability, (c) distance to industrial centers, (d) distance to small roads, and (e) distance to primary roads (Huang et al. 2009).

Figure 8. Neighbor index (left) and constraints (right).

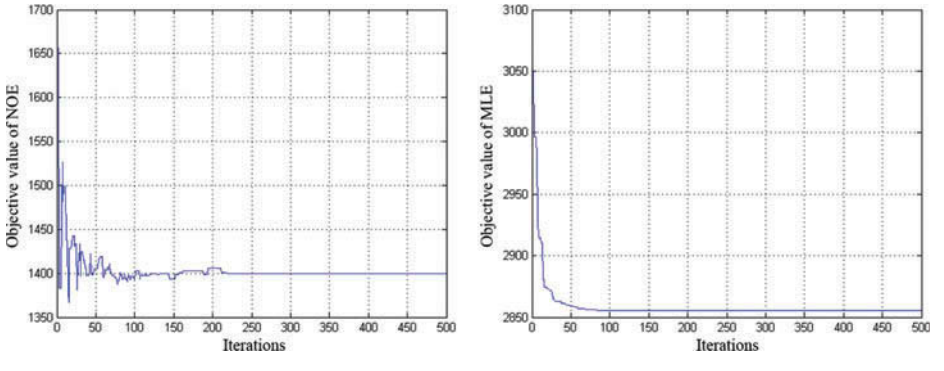

Figure 9. Convergence curves of NOE and MLE values toward the objective of MLE.

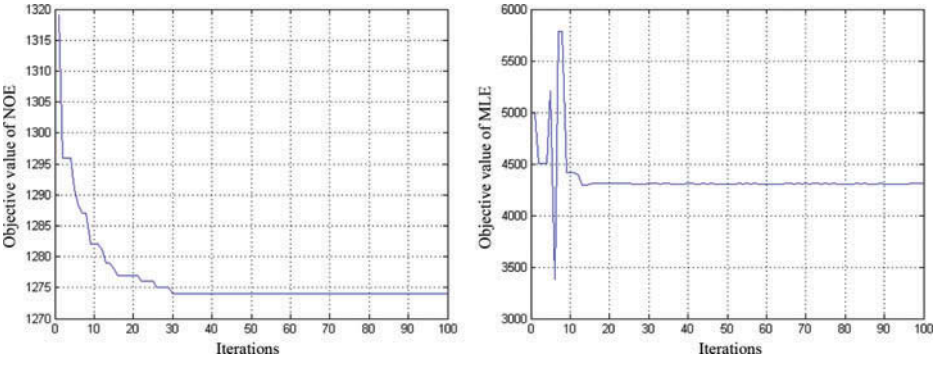

Figure 10. Convergence curves of NOE and MLE values toward the objective of NOE.

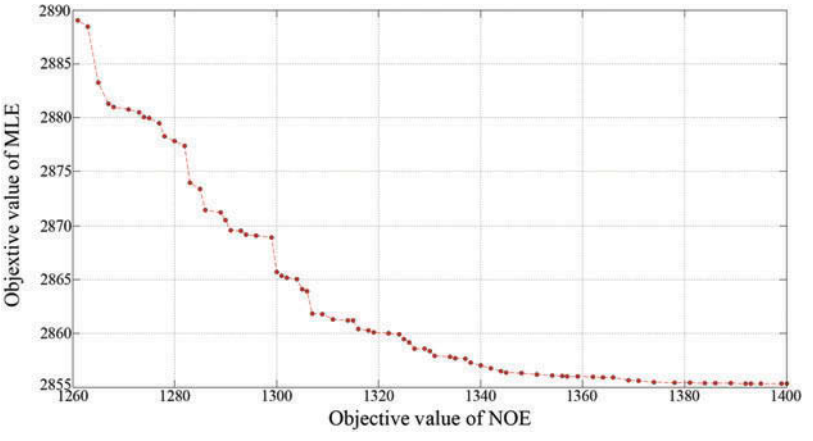

Figure 11. Pareto front of NSGA-II optimization model.

convergence of the optimization process, and the MLE value reaches 2855. However, with respect to the left curve, which could show the changes in the objective value along with the iterations, the NOE value varies rapidly for the first 200 iterations and converges at around 1400 after 200 iterations. This clearly presents the gap or conflict between the two objectives. In this study, the optimization process is repeated 10 times and a unique set of parameters is obtained, which can prove the robustness of the computation. The coefficients and constant are in the first row in Table 3.

Second, the NOE-based GGA calibration approach is used, and the convergence curves are separately for the objectives of NOE and MLE by 100 iterations. In Figure 10, the convergent curve of the NOE and the unordered curve of MLE also prove that there is a gap between the two objectives. In addition, the 10 runs for this optimization also reveal the instability, as it becomes trapped in the local optimum, of the NOE as a single optimization objective, although it can, at times, produce a solution with a better NOE (Table 2). Due to the instability of these coefficients, the solutions produced by this approach might not be suitable for simulation without a specific design.

In Table 2, Coeff_a, Coeff_b, Coeff_c, Coeff_d, and Coeff_e are the $(a_1, a_2, ..., a_n)$ in Equation (4), which are the coefficients of these global factors. The Const is the constant a_0 in Equation (4). NOE is the objective value that could also represent the predication precision of these 5000 samples.

Table 2. Parameters of regression and Pareto front solutions.

	Coeff_a	Coeff_b	Coeff_c	Coeff_d	Coeff_e	Const	NOE
Solution_1	−3.753	−6.806	2.011	−10.398	−3.151	6.312	1265
Solution_2	−2.419	−7.753	−7.903	−7.323	−6.441	8.785	1274
Solution_3	−0.161	−10.226	−9.215	4.075	10.828	7.774	1267
Solution_4	−1.925	−10.742	−9.538	−9.581	7.559	9.065	1265
Solution_5	−0.613	−5.215	−5.258	−8.849	−3.688	5.925	1275
Solution_6	−1.387	−10.677	5.194	−10.419	−8.634	8.462	1260
Solution_7	−0.161	−9.796	−5.516	−7.366	9.151	7.065	1258
Solution_8	−2.441	−9.860	2.183	−10.441	−1.409	9.065	1264
Solution_9	−0.290	−6.032	1.129	−3.516	−2.742	5.602	1266
Solution_10	−6.032	−7.968	−6.527	−8.140	8.054	6.742	1274

Finally, the NSGA-II based calibration model is used for 5000 iterations to reach convergent status, which is proved by the comparison of generations that has almost no difference. On the Pareto front, 100 solutions are distributed from 1200 to 1400 for the NOE objective and from 2855 to 2890 for the MLE objective, which means that the solutions not only reach the similar result through the GLRCA, but also exceed the result reach using the NOE-based GGA approach in the earlier experiments (Figures 9 and 10).

4.4. Comparison of results (calibration parameters and simulation precision) from the Pareto front and generic Logit regression solutions based on the whole data set

Figure 11 indicates that there are 100 solutions on the Pareto front at one time. Each solution should be one of these near-optimal solutions with respect to one set of preferences for the two objectives: MLE and NOE. To validate the meanings of these solutions, three representatives – the solution on the Pareto front up, which is the solution with the best NOE value; that on the Pareto front middle, which is the solution with the most similar preference for the two objectives; and the Pareto front down, which is the solution with the best MLE value – are compared using the parameters computed by the GLRCA based on both the 5000 samples and the whole dataset. The results are listed in Table 3.

Coeff_a, Coeff_b, Coeff_c, Coeff_d, and Coeff_e are the $(a_1, a_2, ..., a_n)$ in Equation (4), which are the coefficients of these global factors. Const is the constant a_0 in Equation (4). The MLE is the objective value of MLE and NOE is the objective value of NOE, which could also represent the predication precision of these 5000 samples. In addition, Sim P is the final simulation precision of the land use from 1992 to 1997 based on the whole data set, which totally has 379,149 land cells and has certain 14,999 land cells changed from rural to urban land use.

With regard to the GLRCA, the coefficients of these factors and the constant are representative. The values clearly reveal the relationships between the five factors and the transition preference value. With decreases in the slope, zoning suitability number, distance to small roads, and distance to primary roads, the transition preference value increases. In addition, the greater the distance to industrial centers, the greater the transition preference value. These relationships accord with common sense, and the knowledge that the factor of distance to industrial centers negatively influences the rural–urban conversion can also be understood in relation to the fact that development is focusing on residential and commercial use. With respect to the other three solutions, there are few differences among them, which could also prove the stability and credibility of the optimization-simulation model. The Pareto front up solution has the best NOE and Sim P, but the worst MLE. This means that both the predication precisions of the 5000 samples and the simulation are better than the GLRCA, with very little sacrifice of MLE value. It's also clear that there is almost no difference between the solution produced by

Table 3. Parameters of regression and selected Pareto front solutions.

	Coeff_a	Coeff_b	Coeff_c	Coeff_d	Coeff_e	Const	MLE	NOE	Sim P
Logit regression	−1.830	−2.820	0.857	−10.575	−1.389	2.279	2855.340	1401	76.17
Pareto front Up	−0.419	−2.720	1.323	−2.634	−2.182	2.161	2889.058	1261	78.03
Pareto front middle	−1.882	−2.828	0.376	−0.505	−0.892	2.161	2865.717	1300	76.51
Pareto front down	−1.860	−2.828	0.849	−10.720	−1.387	2.290	2855.343	1400	76.15

the GLRCA and the Pareto front down solution, which suggests that the essential elements of the GLRCA are the same as the multi-objective optimization-simulation, such that the solutions based on the multi-objective optimization-simulation are credible. Meanwhile, the Pareto front middle solution produces the calibrated parameter with compromising NOE, Sim P, and MLE values. A comparison of these four solutions clearly reveals that the proposed model for the calibration of CA-based rural–urban land-use conversion models is capable of generating better solutions than those produced by the GLRCA. It is also able to deliver more effective solutions with varied preferences for the two objectives at one time, which is a novel and meaningful development in the calibration of the parameters in CA-based rural–urban land conversion models.

With respect to the coefficients and constants themselves, the Pareto front up solution shows better precision with less influence of slope and distance to small roads, similar zoning suitability influence, and more influence exercised by the distance to industrial centers and the distance to primary roads than in the GLRCA. Meanwhile, the Pareto front middle solution displays less influence of the distance to industrial centers, small roads, and primary roads in the rural–urban land-use conversion and better NOE and Sim P as well. As previously mentioned, the Pareto front down solution is almost the same as the GLRCA solution for all these coefficients, the constant, NOE, and Sim P.

In addition, the simulation results of the GLRCA and Pareto front solutions are also presented in Figure 12. As for the comparison of these four simulation results, solution (a) is almost totally the same as solution (d), which could also be observed from Table 3. However, there are some differences among solutions (b), (c), and (d) (same as (a)); through scrutiny, we can find in the left-middle of the map, both solutions (b) and (c) have a more accurate simulated urban area (in a shape of 'boot'), besides, we can also find solution (c) has a better simulated urban area than that of both solutions (a) and (b) in the right-bottom area.

5. Conclusion

The calibration of CA rules remains a challenging but essential step in the modeling of CA-based rural–urban land conversion given the importance of the simulation accuracy (goodness of fit) and reasonability of the parameters (statistical meaning), and conflicts between the these two aspects. This study innovatively implements the NSGA-II model to calibrate the parameters of a CA model for the simulation of rural–urban land conversion. As one type of Pareto front-based heuristic optimization algorithms, NSGA-II, with its powerful elitism, non-dominated sorting method, and crowding distance computation, is used to simultaneously calibrate the parameters of the CA model, considering the two objectives of MLE and NOE. Compared to the GLRCA and the NOE-based GGA calibration approach, the characteristics of the NSGA-II-based calibration model make it capable of generating a number of optimal/near-optimal solutions (sets of parameters) on the Pareto front under the consideration of the two objectives mentioned above, which manages to achieve a balance between the pursuing of the simulation accuracy and the reasonability of the parameters (statistical meaning).

Based on the simulation of rural–urban land conversion in New Castle County, the NSGA-II-based calibration model produces 100 sets of parameters on the Pareto front. Three representative solutions from the Pareto front, located at the top-left, middle, and lower-right, are chosen for the comparison of solutions obtained by the GLRCA, which clearly shows that the NSGA-II-based calibration model not only generates solutions with

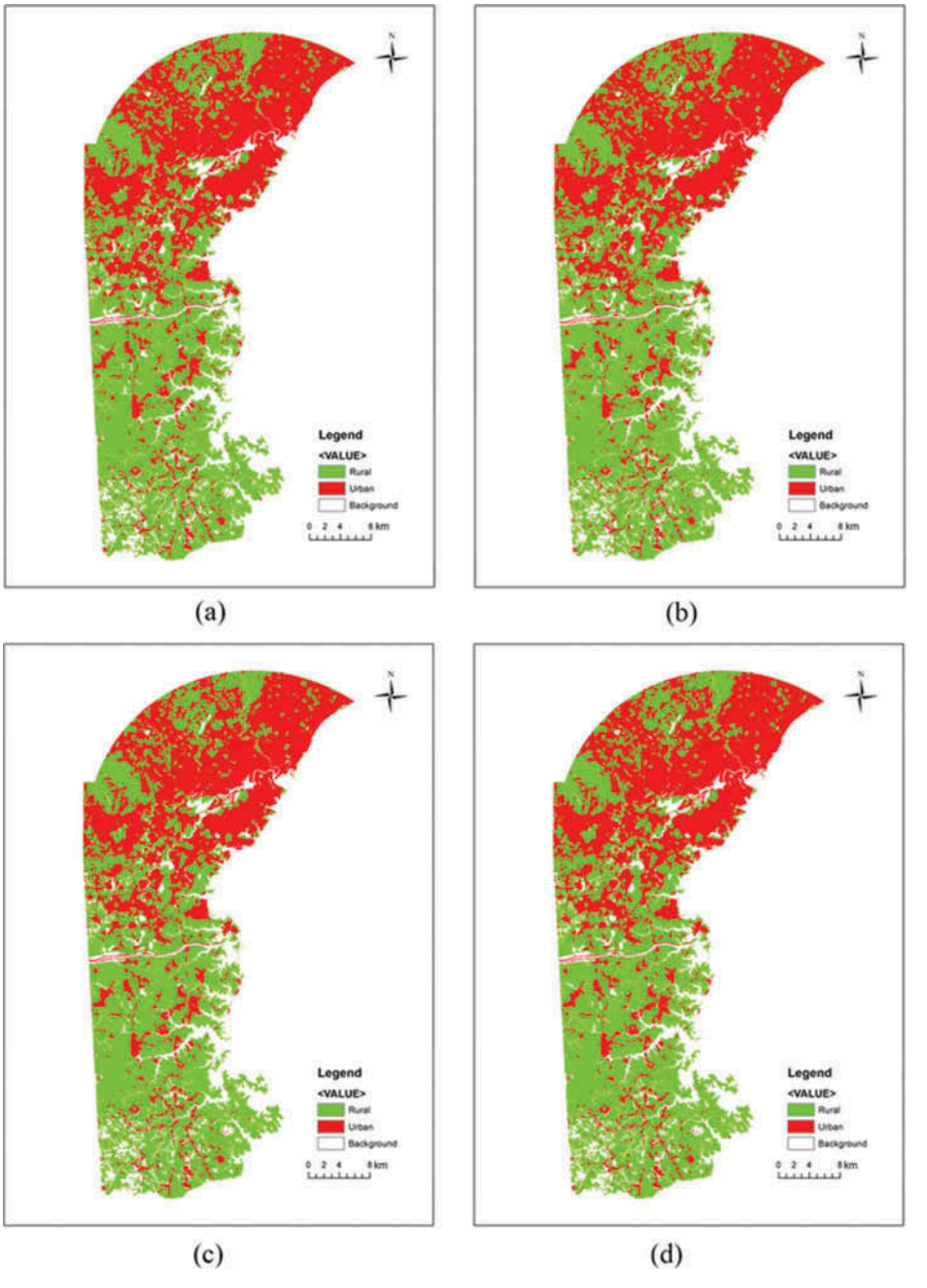

Figure 12. Simulation results of regression and Pareto front solutions. ((a) regression simulation result, (b) Pareto front up simulation result, (c) Pareto front middle simulation result, (d) Pareto front down simulation result).

better precision on the 5000 samples and better simulation result from 1992 to 1997 for the whole research area, but also provides other solutions with better statistical meaning (reasonability). There are more solutions when simulation accuracy and reasonability are

considered under different weights. This could supply the CA model with much more reasonable and correct sets of parameters on the Pareto front.

Given the efficiency of the NSGA-II-based calibration approach, which takes around 30 hours for 5000 iterations, developing its ability to integrate a more efficient algorithm or even parallel computation could be a fruitful direction for future research. Ensuring the generalizability of the calibration model would also be a worthwhile research direction.

Acknowledgments

The supports and comments from Dr. Hongxia Wang, Prof. Yee Leung, Prof. Michael Batty, Prof. Shaowen Wang, and Prof. Michael Goodchild on this research are gratefully acknowledged. We would also like to thank the anonymous reviewers for their valuable comments and suggestions.

References

Batty, M., 1998. Urban evolution on the desktop: simulation with the use of extended cellular automata. *Environment and Planning A*, 30, 1943–1967.

Batty, M. and Torrens, P., 2005. Modeling and prediction in a complex world. *Futures*, 37 (7), 745–766.

Batty, M. and Xie, Y., 1994a. Modelling inside GIS: Part 2. Selecting and calibrating urban models using ARC-INFO. *International Journal of Geographical Information Science*, 8 (5), 451–470.

Batty, M. and Xie, Y., 1994b. From cells to cities. *Environment and Planning B: Planning and Design*, 21, 531–548.

Batty, M., Xie, Y., and Sun, Z., 1999. Modeling urban dynamics through GIS-based cellular automata. *Computers, Environment and Urban Systems*, 23 (1999), 205–233.

Brown, D.G., *et al*., 2005. Path dependance and the validation of agent-based spatial models of land use. *International Journal of Geographical Information Science*, 19 (2), 153–174.

Cao, K., *et al*., 2011. Spatial multi-objective land use optimization: extensions to the nondominated sorting genetic algorithm-II. *International Journal of Geographical Information Science*, 25 (12), 1949–1969.

Clarke, K. and Gaydos, L., 1998. Loose-coupling a cellular automaton model and GIS: long-term urban growth prediction for San Francisco and Washington/Baltimore. *International Journal of Geographical Information Science*, 12 (7), 699–714.

Clarke, K., Hoppen, S., and Gaydos, L., 1997. A self-modifying cellular automaton model of historical urbanization in the San Francisco Bay area. *Environment and Planning B: Planning and Design*, 24 (2), 247–261.

Colonna, A., *et al*., 1998. Learning urban cellular automata in a real world: The case study of Rome metropolitan area. *In: ACRI'98 third conference on cellular automata for research and industry, Trieste, 7–9 October 1998.* London: Springer, 165–18.

Congalton, R.G., 1988. A comparison of sampling schemes used in generating error matrices for assessing the accuracy of maps generated from remotely sensed data. *Photogrammetric Engineering & Remote Sensing*, 54, 593–600.

Deb, K., *et al*., 2000. A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II. *In: 6th international conference on parallel problem solving from nature, Paris, France, 18–20 September, 2000 Proceedings.* Berlin Heidelberg: Springer, 849–858.

Deb, K., *et al*., 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6 (2), 182–197.

Feng, Y. and Liu, Y., 2012. An optimised cellular automata model based on adaptive genetic algorithm for urban growth simulation. In: W. Shi, A. Yeh, Y. Leung & C. Zhou, eds. *Advances in spatial data handling and GIS: 14th international symposium on spatial data handling.* Heidelberg, Germany: Springer, 27–38.

Feng, Y. and Liu, Y., 2013. A heuristic cellular automata approach for modelling urban land-use change based on simulated annealing. *International Journal of Geographical Information Science*, 27 (3), 449–466.

Goldberg, D.E., 1989. *Genetic algorithms in search*. Massachusetts: Addison-Wesley.

Goldstein, N.C., 2003. Brains vs. brawn – Comparative strategies for the calibration of a cellular automata-based urban growth model. *In: 7th international conference on geocomputation. Southampton, UK, 8–10 September 2003 Proceedings*. Boca Raton, FL: CRC Press, 249–272.

Huang, B., Zhang, L., and Wu, B., 2009. Spatiotemporal analysis of rural–urban land conversion. *International Journal of Geographical Information Science*, 23 (3), 21.

Kok, K., *et al*., 2001. A method and application of multi-scale validation in spatial land use models. *Agriculture, Ecosystems & Environment*, 85 (1–3), 223–238.

Landis, J. and Zhang, M., 1998. The second generation of the California urban futures model, Part 2: Specification and calibration results of the land use change submodel. *Environment and Planning B*, 25, 795–842.

Li, X. and Liu, X., 2007. Case-based celular automaton for simulating urban development in a large complex region. *ActaGeographicaSinica*, 62 (10), 1097–1109.

Li, X., Yang, Q., and Liu, X., 2007. Genetic algorithms for determining the parameters of cellular automata in urban simulation. *Science in China Series D: Earth Sciences*, 50, 1857–1866.

Li, X. and Yeh, A., 1999. Constrainted cellular automata for modeling sustainable urban forms. *ActaGeographicaSinica*, 54, 4.

Li, X. and Yeh, A.G.-O., 2000. Modelling sustainable urban development by the integration of constrained cellular automata and GIS. *International Journal of Geographical Information Science*, 14 (2), 131–152.

Li, X. and Yeh, A., 2002. Neural-network-based cellular automata for simulating multiple land use changes using GIS. *International Journal of Geographical Information Science*, 16 (4), 323–342.

Matthews, K.B., (2001). *Applying genetic algorithms to multi-objective land-use planning*. Thesis (PhD). The Robert Gordon University.

Pontius, G.R. and Malanson, J., 2005. Comparison of the structure and accuracy of two land change models. *International Journal of Geographical Information Science*, 19 (2), 243–265. doi:10.1080/13658810410001713434.

Shan, J., Alkheder, S., and Wang, J., 2008. Genetic algorithms for the calibration of celluar automata urban growth modeling. *Photogrammetric Engineering & Remote Sensing*, 74 (10), 1267–1277.

Silva, E.A. and Clarke, K.C., 2002. Calibration of the SLEUTH urban growth model for Lisbon and Porto, Portugal. *Computers, Environment and Urban Systems*, 26 (2002), 525–552.

Vliet, J.V., Bregt, A.K., and Hagen, A., 2011. Revisiting Kappa to account for change in the accuracy assessment of land-use change models. *Ecological Modelling*, 222 (8), 1367–1375.

VonNeumann, J. and Burks, A.W., 1966. *Theory of self-reproducing automata*. Urbana: University of Illinois Press.

White, R. and Engelen, G., 1993. Cellular automata and fractal urban form: a cellular modelling approach to the evolution of urban land-use patterns. *Environment and Planning A*, 25, 1175–1199.

White, R. and Engelen, G., 1997. Cellular automata as the basis of integrated dynamic regional modelling. *Environment and Planning B: Planning and Design*, 24, 235–246.

Wu, F., 1998. SimuLand: a prototype to simulate land conversion through the integrated GIS and CA with AHP-derived transition rules. *International Journal of Geographical Information Science*, 12, 63–82.

Wu, F., 2002. Calibration of stochastic cellular automata: the application to rural–urban land conversions. *International Journal of Geographical Information Science*, 16 (8), 795–818.

Wu, F. and Webster, C.J., 1998. Simulation of land development through the integration of cellular automata and multi-criteria evaluation. *Environment and Planning B: Planning and Design*, 25, 103–126.

Yang, Q., 2008. Dynamic transition rules for geographical cellular automata. *Journal of Sun Yatsen University (Natural Science Edition)*, 47, 4.

Yang, Q. and Li, X., 2007. Calibrating urban cellular automata using genetic algorithms. *Geographical Research (In Chinese)*, 26 (2), 226–231.

Yang, Q., Li, X., and Shi, X., 2007. Cellular automata for simulating land use changes based on support vector machines. *Computers & Geosciences*, 34 (6), 592–602.