MYU Tokyo

S & M 3221

Data Mining of National Geographical Census for Decision-making in Urban Planning: A Geo-simulation of Urban Size in Beijing, China

Miao Wang,[1,2,3] Meizi Yang,[4] Xu-dong Yang,[1,2] Juan Chen,[1,2] and Bogang Yang[1,2]*

[1]Beijing Institute of Surveying and Mapping, Beijing 100038, China
[2]Beijing Key Laboratory of Urban Spatial Information Engineering, Beijing 100038, China
[3]State Key Laboratory of Geo-Information Engineering and Key Laboratory of Surveying and Mapping Science and Geospatial Information Technology of MNR, CASM, Beijing 100036, China
[4]Beijing University of Civil Engineering and Architecture, Beijing 100044, China

(Received November 7, 2022; accepted January 24, 2023)

Keywords: boundary calculation, construction land, cellular automaton, machine learning, integrated learning

With the development of the census and monitoring of national geographical conditions in China, the availability of information has sharply increased. Progress in data mining methods and social application tools has provided a way for solving the problems of low resource allocation and high uncertainty in decision-making regarding planning. To relieve non-capital functions and serve the healthy development of the Beijing Metropolitan Area, we propose a new model of self-adaptive cellular automaton based on ensemble learning (EL-CA). The method is based on the data collected by monitoring geographical conditions and is guided by complex geocomputing that simulates city-scale evolution in Beijing. A comparison of predicted and real data for Beijing in 2015 demonstrated that the predictions made by the EL-CA model proposed significantly outperformed those by traditional cellular automaton (CA) models based on empirical statistics. Data on the geographical conditions in Beijing in 2007 and 2015 were employed in model simulation and training to predict the scale of the city in 2023. The urban agglomeration points in Beijing tended to be dense, the overall construction land tended to be saturated, and the growth rate of land use areas slowed. Results from the model also established that the construction land in Beijing is close to saturation from a quantitative perspective, and the potential urban expansion hotspots in the future are mainly concentrated in the Tongzhou District, the Daxing District, the Fangshan District, the south side of the fourth and fifth ring roads, and the southwest side of Pinggu District. These results can provide decision-makers in urban planning with supporting data and support Beijing to relieve Beijing of functions nonessential to its role as China's capital.

*Corresponding author: e-mail: bogangy@126.com

https://doi.org/10.18494/SAM4241

1. Introduction

National geographical conditions comprehensively express the nature, economy, and human geography at a country's macro level, which is an important part of basic national conditions. A survey of geographical conditions based on remote sensing technology is one of the main aspects of work in monitoring land space in China in recent years. Remote sensing images are an important data source for the survey of geographical conditions. Surveys of geographical conditions use interpretation technology on remote sensing images to achieve rapid extraction and accurate interpretation of geographical information. Monitoring national geographical conditions involves the investigation and observation of status quo and evolution of national geographical information. National conditions and national strength can be accurately and effectively captured by research and statistics based on national geographical information in a certain space and over time. As a subset of national geographical conditions, urban geographical conditions include natural, economic, and human geographic elements, such as urban area and layout, urban topography and land use, population distribution and density, road traffic, and urban expansion. Monitoring of the geographical condition of cities includes two primary aspects: important geographical information and evolutionary geographical information. The former is visualized by carriers such as maps and images, whereas the latter is distributed orderly over time. The direction of future development can be analyzed and predicted by research on how regularly information changes with time.(1–2)

With rapid economic growth, urban population continues to grow, urban scale continues to expand, and urban construction areas continue to increase. This has caused a series of social and ecological problems such as traffic congestion, uneven distribution of social resources, and deterioration of the ecological environment. The evolution of urban scale is affected by many things, including geographical conditions, social economy, infrastructure, population, and national policies and regulations. Therefore, the evolution of urban scale is highly complex. To identify the rules of urban scale evolution, we must first explore the inner driving factors that deeply affect urban scale. After that, we can accurately predict future trends in urban scale development and take effective measures to provide support for the planning and construction of rational and sustainable cities.

The data acquired by the completion of the first census of national geographical conditions provide an opportunity to study the evolution of urban scale over a large time span. We investigated the national geographical information on urban construction land in Beijing. By summarizing the strengths and weaknesses of the former methods of research in urban-scale simulation, urban cellular automaton (CA), a self-adaptive method of CA geographical simulation based on ensemble learning (EL), is proposed. We simulated the dynamic evolution of urban agglomeration in Beijing under the influence of various policies and other unchanging factors and forecast its future development. Our prediction highlights the key areas that need attention to alleviate Beijing urban agglomeration in the coming years. It also provides strong support for Beijing to relieve Beijing of functions nonessential to its role as China's capital and to promote the coordinated development of the Beijing Metropolitan Area.(3)

2. Research Area and Data Sources

Beijing, the capital of China, is the center of politics, economy, culture, education, and international exchange. It is located between 115.7°–117.4° E and 39.4°–41.6° N, with its center at 39°54′20″ N and 116°25′29″ E, covering a total area of 16,410 square kilometers. It is adjacent to Tianjin and surrounded by Hebei Province and Tianjin. For 30 years, both the urbanization of Beijing and the influx of migrants have been accelerating. Before the mode of development changed from "population influx" to "population outflux," the land area in Beijing devoted to construction had been increasing year by year. To simulate Beijing's size, we use the data from the first geographic national conditions and special monitoring of Beijing in 2015. The specific data include:

(1) Raster data: Grade Map Data of Beijing in 2015.

(2) Vector data: Data on construction land in Beijing in 2007, 2011, and 2015; data on road networks in Beijing (including urban roads, expressways, and railway networks); data on the water system (including threadlike rivers, flat-faced lakes, and reservoirs); all levels of the data on the urban center; and data on protected areas with restricted development.

3. Methods

3.1 Geographical simulation

To compensate for deficiencies in the current modeling of geographical information systems, geographical simulation systems have emerged. Their core lies in the establishment of geographical models, and simulation and prediction of complex geographical phenomena are carried out through simulations and other research methods.(4,5) Key technologies of a geographical simulation system include CA and multi-agents. Among these, the cellular automata method is widely used in the modeling of land use and coverage changes due to its high efficiency and simple assumptions. CA is based on the interaction among the cells in local neighborhoods on the grid system, which simulates complex spatial phenomena. Newly generated cell properties are derived from the cells of the previous moment through certain rules of evolution. This bottom-up feature is consistent with the characteristics of urban scale evolution. Therefore, cellular automata are widely used in simulations of urban scale. For example, Li and Ye (2002) studied the expansion of urban land in Dongguan City with good results by combining artificial neural networks and other technical methods in the modeling.(6–12)

A significant disadvantage of the CA approach is that it does not introduce spatial elements and driving forces in the simulation of land use and changes in coverage. To compensate for this shortcoming, some researchers have improved the scientific nature of logical decision-making by combining CA with empirical statistical models, including an ordinary least-squares regression (OLs) model, a spatial autoregressive model, an analytic hierarchy model, a logistic regression model, and a hierarchical multiple regression model. However, as parametric models, these empirical methods also have some drawbacks:

(1) They rely too much on prior knowledge, but there is no consistent and comprehensive knowledge about the processes of land use and changes in coverage.
(2) They cannot process multi-channel data or identify nonlinear correlations. On the other hand, land use and changes in coverage are affected by a series of complex and nonlinear factors.
(3) They cannot process high-dimensional data, which may lead to over-fitting. The existence of these drawbacks greatly reduces the versatility of the models when using the empirical statistical models for out-of-sample prediction. Therefore, it is necessary to find a more reliable and accurate method to replace CA based on the empirical statistical models.(13,14)

Machine learning techniques (such as neural networks, decision trees, random forests, and support vector machines) offer new possibilities for modeling land use and coverage on different geographical scales. Machine learning techniques have the following advantages over empirical statistics:
(1) Stronger processing power for quantitative and qualitative high-dimensional data;
(2) Computational efficiency to avoid overfitting;
(3) No need for strict parameter assumptions to make the description of the data more accurate and reliable;
(4) No strong assumptions are required, and nonlinear features can be processed.

Because of these advantages, some groundbreaking researchers have tried to improve cellular automata by combining different machine learning methods,(15,16) but it is often difficult to choose the “optimal” machine learning algorithm. In addition, there has not been any research up to now on the systematic simulation of urban scale by combining cellular automata and machine learning techniques.

3.2 EL

As an integrated approach, EL methods train multiple machine learning algorithms and integrate their estimates to improve overall numerical prediction or classification accuracy. In general, EL algorithms are superior to single machine learning algorithms because errors change with different machine learning algorithms, but the combination of different machine learning algorithms makes the results more accurate. Typical methods of EL are divided into two types: bagging algorithms and stacking algorithms.(17–23)

The bagging algorithm is a technique to reduce generalization error by combining several models. The main idea is to train several different models, and then let all models vote on the output of test samples. This is an example of conventional strategies in machine learning, also known as model averaging. The bagging algorithm is characterized by training each algorithm in the machine learning algorithm with different subsets of training data, which can effectively reduce the sensitivity of the machine learning algorithm, avoid the occurrence of over-fitting, and improve the overall generalization ability. Specifically, n times of repetitive sampling of variables in the original training samples are first carried out by random sampling to generate n new sample sets. With these new sample sets, m machine learning algorithms are trained separately. Then, the same variables are used as input to these algorithms to obtain m prediction values, and the bagging algorithm finally obtains the integrated output results of the m prediction values by simple voting or averaging methods.

The stacking algorithm is also called the superposition method. First, *m* machine learning algorithms are trained with the existing training sample sets. Second, the output values of the *m* algorithms are used as training samples, and the actual value of the variable is taken as the true value. Finally, the original machine learning algorithm is superimposed onto a new machine learning algorithm through a new training dataset. In the superposition method, the outputs of the previous machine learning algorithm are applied to the induction process of the subsequent algorithm, so that the new machine learning algorithm can identify and correct the errors of the original machine learning algorithm, find the best ensemble method of the machine learning algorithms, and improve the accuracy of learning.

3.3 CA based on self-adaptive EL

We selected five machine learning algorithms commonly used in combination with cellular automata to simulate land use and changes in coverage: support vector machine (SVM), radial basis function neural network (RBF-NN), random forests (RFs), boosted tree regression (BTR), and rough sets (RSs).

SVM is a generalized linear classifier that carries out binary classification of data according to supervised learning. Its decision boundary is the maximum distance hyperplane for learning samples. Finding a hyperplane (segmentation line) to segment the sample segmentation maximizes the interval.

RBF-NN is an artificial neural network that uses a radial basis function as the activation function. It is a three-layer forward network including an input layer, a hidden layer, and an output layer. The transformation from input space to hidden space is nonlinear, while the transformation from hidden space to output space is linear.

RFs refers to a classifier that uses multiple trees to train and predict samples. In fact, it is a special bagging method, which uses the decision tree as the model in bagging.

BTR is a lifting method based on classification trees and regression trees. The decision tree is a binary classification tree for classification problems and a binary regression tree for regression problems.

RSs is a mathematical tool for the quantitative analysis and treatment of imprecise, inconsistent, and incomplete information and knowledge. It is a two-dimensional table that uses information tables to describe objects in the universe. Each row represents an object, and each column represents an attribute of the object. The degree of imprecision is described by the concepts of lower approximation and upper approximation.

On the basis of the data from monitoring Beijing's national geographical conditions in 2007, 2011, and 2015, we extracted the predictive factors of urban construction land and urban scale evolution in Beijing. The predictive factors can be divided into auxiliary factors and control factors, among which the auxiliary factor data includes distance factors (distance from road – subdivided into distance from railway, highway, and expressway), distance from city (town) center, distance to satellite center, distance to rivers, lakes or water systems, and topographic factors (slope data). In view of the many cultural relics and historical protected areas in Beijing, we selected protective areas with restricted development as the limiting factors. The predictive factors and their sources are shown in Table 1.

Table 1
Predictive factors and their sources.

Predictive Factors	Category	Source
Auxiliary Factors	Distance Factors	Distance raster data calculated from vector data such as Beijing road network, water system, and town center.
	Topographic Factors	Slope raster data calculated from Beijing DEM data.
	Other Factors	Other influencing factors based on the results of the census of the geographical conditions of Beijing.
Limiting Factors	Protective Areas with Restricted Development	Converting vector data of the restricted development zones to raster data.

After converting the vector data for Beijing's urban construction land in 2007, 2011, and 2015 into raster data, the CA model based on the five machine learning algorithms described was trained using the self-adaptive approach, and the city scale of Beijing in 2023 was predicted. The data for the first two years were used for model training, the data for the third year were used for adjustment and calibration, and randomness was reduced by repeating 10-fold cross-validations ten times. Specifically, the training dataset was divided into ten subsets, nine of which were used for training, and the last one was used for testing. Then the bagging algorithm and the stacking algorithm were used to generate a new CA model based on EL. These five machine learning algorithms can generate many new superposition algorithms, which can use the Wilcoxon signed-rank test to compare the saliency of the differences shown by different superposition combinations and can adopt the superposition combination with the least number of algorithms and the highest prediction accuracy as the final simulation method. The cellular automata model algorithm was trained with real data from multiple years and was verified and calibrated with one of the years. Randomness was reduced by repeated training and cross validation. Finally, the combination with the highest predictive accuracy was found by the superposition of different algorithms.

In this study, 240 sets of randomly selected samples were used to calculate the predictive accuracy, and the best CA model based on EL was selected. To judge if the model were relatively better than the traditional CA method based on the empirical statistical model, we further compared the accuracy of the two predictions.

4. Results

4.1 Comparison of EL methods and empirical statistical methods

We used various model methods to predict construction land in Beijing in 2015 and compare it with the actual situation in 2015, as shown in Table 2. From the predictions, it can be seen that the accuracy of the machine learning algorithm was higher than that of the empirical statistical model algorithm. At the same time, we also show the accuracy of predictions from the bagging algorithm and stacking algorithm. The accuracy of each machine learning algorithm was significantly improved after being combined with the bagging algorithm, but it was still not as good as that of the stacking algorithm. Among the methods listed, the highest predictive

Table 2
Prediction by each model (%).

Original Model	Prediction Accuracy	Bagging Algorithm	Prediction Accuracy	Stacking Algorithm	Prediction Accuracy
SVM	69.1	SVM	77.5	SVM + RSs	81.1
RBF-NN	65.5	RBF-NN	72.9	SVM + RSs + RBF-NN	84.7
RFs	63.2	RFs	70.5	SVM + RSs + RBF-NN + RFs	74.3
BTR	60.1	BTR	68.4	SVM + RSs + RBF-NN + RFs + BTR	73.9
RSs	68.0	RSs	75.9		
OLs	48.4				
SAR	56.7				

SAR: spatial auto regression

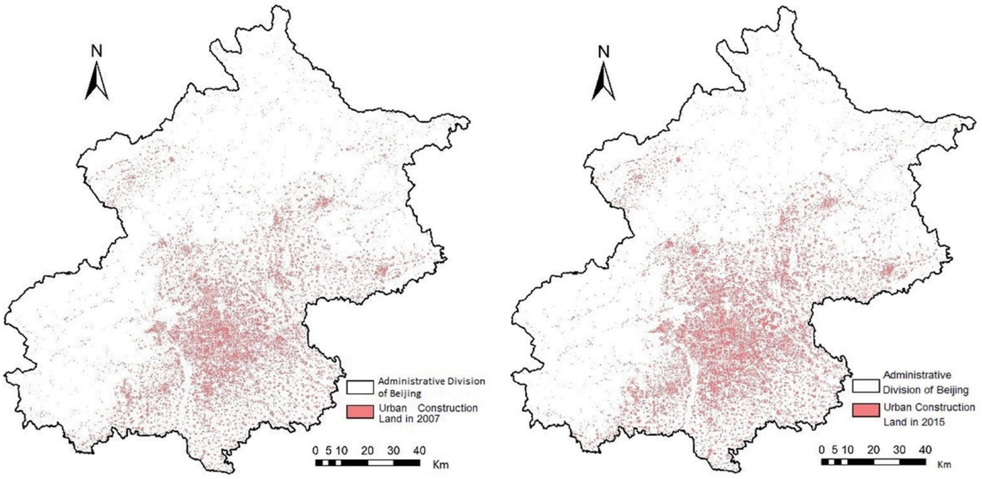
accuracy was achieved when the stacking algorithm superimposed the SVM, rough set, and RBF-NN, for which the prediction accuracy reached 84.7%.

4.2 Predicted Results

The land use for buildings in Beijing in 2007 and 2015 is shown in Fig. 1. Combining this with the results from the predictive models, we selected three algorithms to be superposed upon the stacking algorithms, namely, SVM, RSs, and RBF-NN. On the basis of the data for Beijing's construction land in 2007 and 2015, the urban construction land was predicted for 2023. The comparison of the construction land between the predicted results for 2015 and 2023 is shown in Fig. 2.

From the perspective of spatial distribution, the overall extent of construction land in Beijing has not increased significantly. The internal distribution tends to be dense, while the overall contour is unchanged. The newly added construction land is mainly distributed on the edge of existing construction land. After excluding the influence of random points, there is no obvious new point within the fourth ring road of Beijing, which shows that construction land in the downtown area of Beijing has become saturated. New sites are primarily concentrated outside the fourth and fifth ring roads, such as on the south side of the Tongzhou District, the Daxing District, and the Fangshan District, and the southwest side of the Pinggu District, all of which are close to Tianjin City and Hebei Province, as the key areas for relieving Beijing of functions nonessential to its role as the capital. The construction land of the Huairou District, the Changping District, the Mentougou District, and the Fangshan District remains mainly unchanged, which may be related to steep slopes of these areas and a large number of areas in which development was prohibited.

Urban construction land predicted in 2007, 2015, and 2023 was converted into raster data with a cell size of 30 × 30. From the absolute value of the change in the area, there were 267539 newly added pixels in 2015 as compared to 2007, and 259854 predicted newly added pixels in


N
Administrative Division of Beijing
Urban Construction Land in 2007
0 5 10 20 30 40
Km
N
Administrative Division of Beijing
Urban Construction Land in 2015
0 5 10 20 30 40
Km

(a) (b)

Fig. 1. (Color) Urban construction land in Beijing in (a) 2007 and (b) 2015.

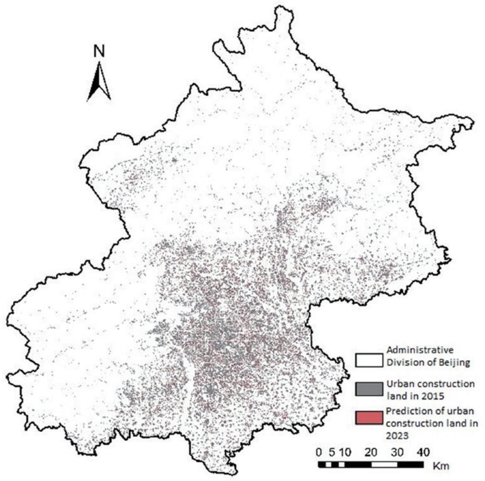

N
Administrative Division of Beijing
Urban construction land in 2015
Prediction of urban construction land in 2023
0 5 10 20 30 40
Km

Fig. 2. (Color) Comparison of predicted results with actual urban construction land in 2015.

2023 compared to 2015. The results show that, if policy and other factors remain unchanged, the growth rate of construction land in Beijing decreased with the increase in area, which also indicates that construction land in Beijing is close to saturation due to slower growth in the overcrowded city. In the future of urban construction and planning, Beijing should strictly control the quantity of new construction land, improve road traffic conditions in the city, alleviate the uneven distribution of population, and protect the environment.

On the basis of the results from data on urban construction land in 2007, 2011, and 2015 in the general survey of geographical conditions, we verify the use of remote sensing technology in the general survey of geographical conditions and that it provides support for decision makers on urban planning. Remote sensing technology not only greatly improves the speed of the census of geographical conditions but also improves the accuracy of the census.

5. Conclusions

To study the evolution of Beijing's urban scale and provide decision support for Beijing's policy of alleviating non-capital functions and promoting the coordinated development of Beijing, Tianjin, and the Hebei Region, we summarized the experiences of predecessors using CA to simulate the urban scale, analyzed the advantages and disadvantages of various methods, and emphasized a CA model based on a machine learning algorithm and EL. Through the examples in this paper, we verified that the method proposed is much better than the traditional modeling method based on empirical statistics in terms of the predictive accuracy of urban scale simulation.

On the basis of the survey data of Beijing's geographical conditions in 2007 and 2011 as the input data, the 2015 predicted value was obtained. Comparing it with the actual value in 2015, we selected the CA model with the highest predictive accuracy; it consisted of a stacking algorithm superimposing SVM, RSs, and RBF-NN. Finally, the distance factor, slope factor, limiting factor, and other factors were provided as input. According to the data of Beijing's construction land in 2007 and 2015, the scale of construction land in Beijing in 2023 was predicted under the condition of unchangeable external factors, and the city scale of Beijing was analyzed.

The results show that the overall contour of Beijing's construction land is basically unchanged, the internal agglomeration point tends to be dense, the overall construction land tends to be saturated, and the growth rate of land area has slowed. To achieve orderly and sustainable urban development, the government needs to relieve Beijing of functions nonessential to its role as the capital. In addition to the central urban area, key control areas for future construction land include the south side of the Tongzhou District, the Daxing District, and the Fangshan District, and the southwest side of the Pinggu District, all of which are close to Tianjin and Hebei Province.

The experimental results in this paper fully demonstrate the flexibility and predictive accuracy of the CA model based on the EL method in urban scale simulation, but there are also deficiencies. In addition to the impact of the extracted impact factors from the monitoring of national geographical conditions, the city scale is also affected to a certain extent by policy formulation and other aspects. To further improve the accuracy of urban prediction, further information needs to be obtained to provide more accurate references for decision makers.

Acknowledgments

The authors thank the State Key Laboratory of Geo-Information Engineering and the Key Laboratory of Surveying and Mapping Science and Geospatial Information Technology of MNR, CASM (No. 2022-04-05) and the Beijing Key Laboratory of Urban Spatial Information Engineering (No. 20210102) for financial support for this study.

References

1 D. Li, H. Sui, and J. Shan: Geomatics Inf. Sci. Wuhan Univ. **37** (2012) 505. https://doi.org/10.13203/j.whugis2012.05.011

2 F. Zhao: Shanghai Land Resour. **32** (2011) 74. https://doi.org/10.3969/j.issn.2095-1329.2011.03.020

3 B. Yang and X. Yu: Sci. Surv. Mapp. **39** (2014) 47. https://doi.org/10.16251/j.cnki.1009-2307.2014.12.013

4 X. Li, J. Ye, and X. Liu: City Plann. Rev. **30** (2006) 69. https://doi.org/10.3321/j.issn:1002-1329.2006.06.015

5 X. Li: Geographical Simulation Systems: Cellular Automata and Multi-Agent Systems (Science Press, Beijing, 2007) 1st ed., p. 250.

6 X. Guo and S. Tang: J. Geomatics **46** (2021) 96. https://doi.org/10.14188/j.2095-6045.2019093.

7 S. Gao: Eng. Tech. Res. **5** (2020) 246. https://doi.org/10.19537/j.cnki.2096-2789.2020.15.115.

8 S. Wang: Geomatics Spatial Inf. Technol. **43** (2020) 41. https://doi.org/CNKI:SUN:DBCH.0.2020-05-013

9 Z. Liang, J. Zuo, and X. Zhu: Geomatics Spatial Inf. Technol. **42** (2019) 56. https://doi.org/CNKI:SUN:DBCH.0.2019-01-016

10 Z. Wang, Q. Zhan, S. Tang, and J. Liu: J. Geomatics. **44** (2019) 23. https://doi.org/10.14188/j.2095-6045.2018409.

11 L. Wang, Y. Luo, and H. Jiao: Urbanism Archit. **1** (2015) 325. https://doi.org/10.3969/j.issn.1673-0232.2015.29.296

12 X. Li and J. Ye: Acta Geogr. Sin. **57** (2002) 159. https://doi.org/10.3321/j.issn:0375-5444.2002.02.005

13 S. Su, Y. Sun, C. Lei, M. Weng, and Z. Cai: Land Use Policy **67** (2017) 415. https://doi.org/10.1016/j.landusepol.2017.06.011

14 K. Kang: Shandong Norm. Univ. (2009). https://doi.org/CNKI:CDMD:2.2009.126064

15 R.-M. Basse, O. Charif, and K. Bódis: Appl. Geogr. **67** (2016) 94. https://doi.org/10.1016/j.apgeog.2015.12.001

16 F. Wang, J.-G. Hasbani, X. Wang, and D. J. Marceau: Comput. Environ. Urban Syst. **35** (2011) 116. https://doi.org/10.1016/j.compenvurbsys.2010.10.003

17 J. Demšar: J. Mach. Learn. Res. **7** (2006) 1. https://jmlr.org/papers/volume7/demsar06a/demsar06a.pd

18 Z. Zhou: Ensemble Methods: Foundations and Algorithms (Electronic Industry Press, Beijing, 2020) 1st ed., Chap. 3.

19 A. Mokhtari, B. Tashayo, and K. Deilami: Int. J. Environ. Res. Public Health **18** (2021) 7115. https://doi.org/10.3390/IJERPH18137115

20 G. Luo, M. Wu, and Z. Pang: J. Syst. Sci. Complexity **34** (2021) 2310. https://doi.org/10.1007/S11424-021-1088-Y

21 Y. Zheng, L. Gao, S. Li, and D. Wang: Energy **239** (2022) 122033. https://doi.org/10.1016/J.ENERGY.2021.122033

22 A. Palazón-Bru, M. I. Tomás-Rodríguez, M. T. López-Cascales, D. M. Folgado-de la Rosa, and V. F. Gil-Guillén: J. Pediatr. Adolesc. Gynecology **30** (2017) 664. https://doi.org/10.1016/j.jpag.2017.06.006

23 R. Anders, Z. Oravecz, and F.-X. Alario: Behav. Res. Methods **50** (2018) 989. https://doi.org/10.3758/s13428-017-0921-7