

Data Analysis and Optimization for Intelligent Transportation in Internet of Things

Lead Guest Editor: Naixue Xiong

Guest Editors: Hongju Cheng, Sajid Hussain, and Changhoon Lee





Data Analysis and Optimization for Intelligent Transportation in Internet of Things

Journal of Advanced Transportation

**Data Analysis and Optimization for
Intelligent Transportation in Internet of
Things**

Lead Guest Editor: Naixue Xiong





Guest Editors: Hongju Cheng, Sajid Hussain, and
Changhoon Lee



Copyright © 2021 Hindawi Limited. All rights reserved.























This is a special issue published in "Journal of Advanced Transportation." All articles are open access articles distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Associate Editors

Juan C. Cano , Spain
Steven I. Chien , USA
Antonio Comi , Italy
Zhi-Chun Li, China
Jinjun Tang , China

Academic Editors

Kun An, China
Shriniwas Arkatkar, India
José M. Armingol , Spain
Socrates Basbas , Greece
Francesco Bella , Italy
Abdelaziz Bensrhair, France
Hui Bi, China
María Calderon, Spain
Tiziana Campisi , Italy
Giulio E. Cantarella , Italy
Maria Castro , Spain
Mei Chen , USA
Maria Vittoria Corazza , Italy
Andrea D'Ariano, Italy
Stefano De Luca , Italy
Rocío De Oña , Spain
Luigi Dell'Olio , Spain
Cédric Demonceaux , France
Sunder Lall Dhingra, India
Roberta Di Pace , Italy
Dilum Dissanayake , United Kingdom
Jing Dong , USA
Yuchuan Du , China
Juan-Antonio Escareno, France
Domokos Esztergár-Kiss , Hungary
Saber Fallah , United Kingdom
Gianfranco Fancello , Italy
Zhixiang Fang , China
Francesco Galante , Italy
Yuan Gao , China
Laura Garach, Spain
Indrajit Ghosh , India
Rosa G. González-Ramírez, Chile
Ren-Yong Guo , China





Yanyong Guo , China
Jérôme Ha#rri, France
Hocine Imine, France
Umar Iqbal , Canada
Rui Jiang , China
Peter J. Jin, USA
Sheng Jin , China
Victor L. Knoop , The Netherlands
Eduardo Lalla , The Netherlands
Michela Le Pira , Italy
Jaeyoung Lee , USA
Seungjae Lee, Republic of Korea
Ruimin Li , China
Zhenning Li , China
Christian Liebchen , Germany
Tao Liu, China
Chung-Cheng Lu , Taiwan
Filomena Mauriello , Italy
Luis Miranda-Moreno, Canada
Rakesh Mishra, United Kingdom
Tomio Miwa , Japan
Andrea Monteriù , Italy
Sara Moridpour , Australia
Giuseppe Musolino , Italy
Jose E. Naranjo , Spain
Mehdi Nourinejad , Canada
Eneko Osaba , Spain
Dongjoo Park , Republic of Korea
Luca Pugi , Italy
Alessandro Severino , Italy
Nirajan Shiwakoti , Australia
Michele D. Simoni, Sweden
Ziqi Song , USA
Amanda Stathopoulos , USA
Daxin Tian , China
Alejandro Tirachini, Chile
Long Truong , Australia
Avinash Unnikrishnan , USA
Pascal Vasseur , France
Antonino Vitetta , Italy
S. Travis Waller, Australia
Bohui Wang, China
Jianbin Xin , China



Hongtai Yang , China
Vincent F. Yu , Taiwan
Mustafa Zeybek, Turkey
Jing Zhao, China
Ming Zhong , China
Yajie Zou , China

Contents





Research on Multifeature-Based Superposter Identification in Online Learning Forums

Changri Luo , Xinhua Zhang , Tingting He, Yong Zhang , Neal Xiong, and Zizhou Lu 
Research Article (10 pages), Article ID 1496321, Volume 2021 (2021)


Forecast and Analysis of Coal Traffic in Daqin Railway Based on the SARIMA-Markov Model

Cheng Zhang  and Shouchen Liu 
Research Article (10 pages), Article ID 1276305, Volume 2020 (2020)

Effective Evolutionary Algorithm for Solving the Real-Resource-Constrained Scheduling Problem

Huu Dang Quoc , Loc Nguyen The , Cuong Nguyen Doan , and Naixue Xiong 
Research Article (11 pages), Article ID 8897710, Volume 2020 (2020)

Urban Traffic Flow Forecast Based on FastGCRNN

Ya Zhang, Mingming Lu , and Haifeng Li
Research Article (9 pages), Article ID 8859538, Volume 2020 (2020)

Research on Coordinated Development of a Railway Freight Collection and Distribution System Based on an “Entropy-TOPSIS Coupling Development Degree Model” Integrated with Machine Learning

Yun Jing, Si-Ye Guo , Xuan Wang , and Fang-Qiu Chen 
Research Article (14 pages), Article ID 8885808, Volume 2020 (2020)

Research on the Simulation Application of Data Mining in Urban Spatial Structure

Jun Zhang, Xin Sui, and Xiong He 
Research Article (9 pages), Article ID 8863363, Volume 2020 (2020)

Research Article

Research on Multifeature-Based Superposter Identification in Online Learning Forums

Changri Luo ¹, Xinhua Zhang ², Tingting He,³ Yong Zhang ³, Neal Xiong,⁴
and Zizhou Lu ¹

¹School of Vocational and Continuing Education, Central China Normal University, Wuhan 430079, China

²School of Computer Science Wuhan Vocational College of Software and Engineering, Wuhan 430205, China

³Academy of Computer Science, Central China Normal University, Wuhan 430079, China

⁴Northeastern State University, Department of Mathematics and Computer Science, Talequah, OK, USA

Correspondence should be addressed to Yong Zhang; 12081207@qq.com

Received 22 February 2020; Accepted 3 April 2021; Published 15 April 2021

Academic Editor: Hongju Cheng

Copyright © 2021 Changri Luo et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

With the development of online learning and distance education, online learners' discussions in forums become increasingly effective to facilitate learning. Superposters, who play a more and more important role in forums, have attracted researchers' close attention. The key to the research is how to identify superposters among a large number of participants. Some studies focus on the network interaction of superposters and some content-related features but neglect the basic quality like language expression that a superposter should possess and the learning-related features like learning collaboration. Based on the analysis of online learning corpus, through network interaction and combination of the different features of N-gram, the paper proposed the superposter identification method based on the three primary features including language expression (L), content quality (C), and social network interaction (S) and the eight secondary features including learning collaboration. The paper applied the method in the real online learning forum corpus for identifying 28 preset superposters, achieving the results of $P@15 = 1.0$, $\text{Avg.P@15} = 1.0$, $P@28 = 0.86$, and $\text{Avg.P@28} = 0.95$. Experiments showed that this was an effective superposter identification method in online learning forums.

1. Introduction

With the improvement in online learning and remote education, discussions in online forums become increasingly effective to facilitate learning. Through online messages with teachers, learners may solve problems and ease emotional loneliness during learning. Previous research study has proved that the opinion of leaders plays an important role in online learning and has a positive effect on interactions [1, 2]. Their posts may significantly help themselves and others to learn. To differentiate between the opinion leaders in social networks, this paper terms them as superposters in learning forums. At present, research is almost nonexistent on superposter identification in online learning forums, unlike the opinion leaders in a traditional sense where much research exists [3–10]. In the context of online learning

forums, superposters refer to the users who are active in posting high-quality information, which may help learners to solve problems and prompt learning [11]. Considering the differences in discourse environment, the superposters of online learning forums differ from popular opinion leaders in social networks. Opinion leaders in social networks mainly spread information via the Internet and thus exert an influence on information receivers in terms of public opinions and tend to affect public opinions. Therefore, according to the explanation of superposters and opinion leaders, there are similarities and differences between them. Both are active in interactions; superposters aim to boost cooperative study, but opinion leaders try to influence public opinions by swaying others. How do we identify superposters among thousands of online learners? Previous similar research was made on the basis of social online

communities and applied in the fields of society and economy, but little was based on the forums of online learning platforms and applied in the field of education [5]. Although Reppel [12] asserted the application was applied in education, the research was mainly made on blogs for identifying opinion leaders in online learning communities.

Through analysis of authentic online learning forums and the characteristics of superposters, this paper obtained three points with which superposters and ordinary learners were distinguished, whereby a model framework for superposter identification in online learning forums was constructed. The framework considers both of the network interaction structure of learners and the discourse features of posts, so as to better identify superposters. Experiments showed that considering the above appropriate different features was effective in identifying superposters in online learning forums.

There are two main contributions of this work: one is proposing a new framework to identify the superposters in learning forums, and the other is proving the framework is useful for identifying the superposters in online learning forum, by experimenting on real-online learning forum corpus.

In the following, the related work will be reviewed in Section 2, the superposter identification framework will be detailed in Section 3, the experimental design and result will be analyzed in Section 4, and discussion and summary will be made in full in Section 5.

2. Related Work

Opinion leaders play a significant role in social networks. As a result, identifying opinion leaders in the context of social networks attracts the great attention of the related researchers like those in the fields of sociology and business. The role includes participation in social politics [13], promotion and popularization of new products or services in the field of business, and effect on decisions made by other consumers [6, 12]. According to the current literature, the following are main methods of identifying opinion leaders:

- (1) Identification based on network interaction structure: on the basis of the structure, in combination with users' social influence and attributes of web links, this is to reflect users' centrality and prestige in social networks through web link addresses, such as the famous PageRank, HITS algorithm, and social network analysis, which are used to identify opinion leaders [4, 6, 10, 14]. With these methods, the network interaction structure with graph models is simulated to observe the importance of user nodes, which emphasizes the structure but fails to consider the comments of opinion leaders; moreover, in the event network nodes increase for the purpose of increasing the amount of information, the graph structure will become so complicated that opinion leaders cannot be identified effectively [15].
- (2) Identification in combination with network interaction and post contents: in consideration of such limitations as sole dependence on network interaction, plenty of research is made to identify opinion

leaders in combination with post contents and network interaction. Based on social network analysis and user comments, Bodendorf and Kaiser explored the opinion leaders in online communities and the propagation trend of the public opinions they make [16]. In combination with the features of network structure and user behavior and the emotional features of posts, through analysis of multidimensional features, Cao et al. studied the social network-based opinion leaders [17]. Li and Du constructed an opinion leader identification framework with blog contents, author attributes, reader attributes, and the network relationship between blog authors and readers to identify the opinion leaders committed to word-of-mouth marketing in online social blogs [5]. Although good results were achieved, the above research depended too much on influence or centrality, making it impossible to reflect the quality of the contents published by opinion leaders and thus accurately identify opinion leaders. Meanwhile, they were made based on social networks instead of identification in the field of education. In accordance with the features of expertise, novelty, influence, activity, longevity, and centrality, Li and Ma et al. built an indicator framework to identify opinion leaders [18]. Huang et al. identified superposters according to the quantity and quality of learners' posts in course forums [19]. This is rare in terms of opinion leader (superposters) identification in online learning forums but fails to reflect the quality and role in cooperative study of superposters' posts.

In the opinion of the author, superposters in learning forums are different from opinion leaders in social network, and they must have a certain cultural quality and cooperative study skills, which are not reflected in the above research studies. Therefore, in consideration of the limitations of the abovementioned studies on opinion leader identification, this paper proposes a superposter identification framework based on language expression, content quality, and interaction structure, so as to identify superposters among the participating learners and learning supporters.

3. Superposter Identification Framework

The authors consider that the superposters in online learning forums should be as follows: (1) be active in posting/replying; (2) be excellent in language expression; (3) post high-quality posts and have a good ability to learn, or be knowledgeable, or accurately reflect learning needs, or provide other assistances to online learners. These not only reflect the importance of poster nodes in interaction through forums but also indicate the authority of their posts. Based on these features, the paper proposes the framework (see Table 1) for superposter identification in online learning forums (Chinese as the working language), as shown in Table 1. According to the definition given by this paper, for a superposter, we expect to reflect the language expression level of learners, quality of post contents, and activity of interaction, respectively, through language expression, content quality, and social network interaction.

TABLE 1: Indexes and description of model features.

Items	Features (symbol)	Description
Language expression (Le_Index)	Word normalization (W_I)	To survey the use of Class I and Class II Chinese characters in posts
	Term nonnormalization (T_I)	To survey the use of uncivil words and Internet slangs in posts
	Language elegance (Le_I)	To survey the use of words, phrases, and idioms
Content quality (Cq_Index)	Learning collaboration (Lc_I)	To survey posters' ability to solve others' problems
	Correlation with the thread (Ct_I)	To survey the similarity between the reply content and the thread title
Social network interaction (Sns_Index)	Expertise of content (Ec_I)	To survey the knowledge points involved in posts
	Out-degree centrality (DC_o)	To survey the intermediate status of posters in network structure
	In-degree centrality (DC_I)	To survey the authority of posters in network structure

3.1. Social Network Interaction. In social network analyses, degree centrality is an important index that measures the social interaction of individuals as well as a common index that evaluates the social status and prestige of individuals, including out-degree centrality and in-degree centrality; out-degree centrality is used to reflect the replies of a poster (learner or learning supporter) a_i to others' posts, as expressed with the following formula:

$$DC_o(a_i) = \frac{\text{NumReplyOut}(a_i)}{N-1} \times \frac{\sum_{j=1, j \neq i}^N a_{ij}}{\sum_{i=1}^N \sum_{j=1}^N a_{ij} - \sum_{i=1}^N a_{ii}}, \quad (1)$$

where a_i is the i^{th} learner of learners set A ; N is the total number of learners, similarly hereinafter; $\text{NumReplyOut}(a_i)$ is the reply of a_i to others, i.e., the number of linkout of node a_i in interactive networks, which reflects the importance of node position; $\sum_{j=1, j \neq i}^N a_{ij} / (\sum_{i=1}^N \sum_{j=1}^N a_{ij} - \sum_{i=1}^N a_{ii})$ is the ratio of the number of a_i 's replies to others' posts to the total number of replies (excluding all self-replies), which reflects the degree of interaction in which a_i participates; and traditional algorithms only consider $\text{NumReplyOut}(a_i) / (N-1)$ but ignores the degree of interaction reflected by $\sum_{j=1, j \neq i}^N a_{ij} / (\sum_{i=1}^N \sum_{j=1}^N a_{ij} - \sum_{i=1}^N a_{ii})$.

Degree centrality, also known as Prestige [20], may reflect the replies of other posters to the posts of a_i , as expressed with the following formula:

$$DC_I(a_i) = \frac{\text{NumReplyIn}(a_i)}{N-1} \times \frac{\sum_{j=1, j \neq i}^N a_{ji}}{\sum_{j=1}^N \sum_{i=1}^N a_{ji} - \sum_{i=1}^N a_{ii}}, \quad (2)$$

where a_i is the i^{th} learner of set A ; N is the total number of learners; $\text{NumReplyIn}(a_i)$ is the number of link in of node a_i in interactive networks, which reflects the node prestige; and $\sum_{j=1, j \neq i}^N a_{ji} / (\sum_{j=1}^N \sum_{i=1}^N a_{ji} - \sum_{i=1}^N a_{ii})$ is the ratio of the total

number of others' replies to a post of a_i to the total number of others' replies (excluding all self-replies), which reflects the centrality of a_i in interactive networks but is rarely considered in traditional algorithms.

Therefore, the index of social network interaction of a_i is calculated as follows:

$$\text{Sns_Index}(a_i) = \sigma \times DC_o(a_i) + (1 - \sigma) \times DC_I(a_i), \quad (3)$$

where σ is a weighting parameter.

3.2. Language Expression. Plenty of research on identification of opinion leaders failed to consider the language expression skill of an opinion leader. However, whether in terms of interaction in social networks or online learning forums, an opinion leader or a superposter must ensure fluent language expression and owns a certain cultural quality. If a post involves violent words or unclear expressions all along, no matter how innovative or important it is, other users (learners of learning forums) may refuse to discuss further. For this reason, this paper makes a survey on language expression with three indexes including "word normalization," "term nonnormalization," and "language elegance." These are relatively easily achieved and may reflect the language expression skill of posters.

3.2.1. Word Normalization. Word normalization is to survey the frequency of Class I and Class II commonly used Chinese characters in posts and thus verify the normalization of the words used by learners. When uncommon words are used in posts to appear intellectual, learners may find it difficult to achieve optimal learning, thus limiting the spread of information. To facilitate survey, the index of the normalization of the words used by a_i is defined as follows:

$$W_I(a_i) = 0.9 \times \frac{\text{CH_I_Freq}(a_i)}{\text{Total_CH_Freq}(a_i)} \times \frac{\text{CH_I_Type}(a_i)}{2500} + 0.1 \times \frac{\text{CH_II_Freq}(a_i)}{\text{Total_CH_Freq}(a_i)} \times \frac{\text{CH_II_Type}(a_i)}{1000}, \quad (4)$$

where $\text{CH_I_Freq}(a_i)$ and $\text{CH_II_Freq}(a_i)$, respectively, are the frequency of Class I and Class II commonly used Chinese characters in all posts of a_i ; $\text{Total_CH_Freq}(a_i)$ is the total

frequency of Chinese characters in all posts of a_i ; and $\text{CH_I_Type}(a_i)$ and $\text{CH_II_Type}(a_i)$, respectively, are the number of the types of Class I and Class II Chinese

characters in all posts of a_i . Constants 2500 and 1000, respectively, are the number of the types of Class I and Class II Chinese characters; 0.9 and 0.1 are weighting parameters and empirical values.

3.2.2. Term Nonnormalization. Term nonnormalization is to survey the use of uncivil words by learners (Internet users). Such usage involves impolite, violent, and vulgar words and some Internet slangs in the process of exchange in forums. Thus, to further analyze the normalization of the words use by surveying the use of uncivil words and Internet slangs, the paper defines the index of term nonnormalization of a_i as follows:

$$T_I(a_i) = \frac{C \times \text{UnC_W_Freq}(a_i) + \text{Net_W_Freq}(a_i)}{\text{TotalWFreq}(a_i)}, \quad (5)$$

where $C = 1.2$ as a constant and reflects that uncivil words are more improper than Internet slangs and $\text{UnC_W_Freq}(a_i)$, $\text{Net_W_Freq}(a_i)$, and $\text{TotalWFreq}(a_i)$, respectively, are the frequency of uncivil words and frequency of Internet slangs in all posts of a_i and the total frequency of words.

3.2.3. Language Elegance. Language elegance is to survey the use of fixed phrases (including fixed terms, phrases, and idioms) in posts. Although language derives from life, we cannot deny the fact that “individualized teaching” (yin cai shi jiao-因材施教, Chinese idiom) is more concise, refined, and elegant than “adopting different teaching methods for different students” in terms of expression. If similar expressions are frequently used in a post, we may see the vocabulary and language mastery of the poster. Accordingly, the paper observes the language expression ability based on this. The language elegance of a_i is calculated as follows:

$$\text{Le}_I(a_i) = \sigma_1 \times \frac{\text{CW 1 Freq}(a_i)}{\text{Total W Freq}(a_i)} \times \frac{\text{CW 1 Type}(a_i)}{\text{Type 1 Num}} + \sigma_2 \times \frac{\text{CW 2 Freq}(a_i)}{\text{Total W Freq}(a_i)} \times \frac{\text{CW 2 Type}(a_i)}{\text{Type 2 Num}} + (1 - \sigma_1 - \sigma_2) \times \frac{\text{Idioms Freq}(a_i)}{\text{Total W Freq}(a_i)} \times \frac{\text{Idioms Type}(a_i)}{\text{Type I Num}}, \quad (6)$$

where $\text{CW1 Freq}(a_i)$, $\text{CW2 Freq}(a_i)$, $\text{Idioms Freq}(a_i)$, and $\text{Total W Freq}(a_i)$ are the frequency of commonly used class I words, frequency of commonly used class II words, frequency of idioms in all posts of a_i , and the total frequency of words, respectively; $\text{CW1Type}(a_i)$, $\text{CW2Type}(a_i)$, and $\text{Idioms Type}(a_i)$, respectively, is the number of the types of Class I and Class II words and idioms;

Type 1 Num , Type 2 Num , and Type I Num , respectively, is the total number of the types of Class I and Class II words and idioms; and $\sigma_1 = 0.25$ and $\sigma_2 = 0.35$ as constants, which are the coefficients from locally optimal solutions obtained through repeated experiments and empirical values.

Therefore, the index of language expression of a_i is calculated as follows:

$$\text{Le_Index}(a_i) = \vartheta_1 \times W_I(a_i) - \vartheta_2 \times T_I(a_i) + (1 - \vartheta_1 - \vartheta_2) \times \text{Le}_I(a_i), \quad (7)$$

where ϑ_j as weighting parameters, $j = 1, 2$.

3.3. Content Quality. The content quality of posts directly affects the result of interaction in online learning forums. Therefore, in the process of identifying superposters, surveying the content quality is very important. In surveying the quality of the contents posted by superposters, the paper mainly focuses on three questions:

- (i) Q1: do the post contents help others to solve learning problems?
- (ii) Q2: are replies relevant to a topic?
- (iii) Q3: how about the conformity of post contents with knowledge points?

That is, if a_i is a superposter, his/her posts will be considered high-quality, helpful to others in the process of interaction and highly relevant to a topic (rather than spam or meaningless posts) and to have highly professional knowledge points. Therefore, the paper evaluates content

quality based on the learning collaboration, correlation with the thread, and expertise of content.

3.3.1. Learning Collaboration. Learning collaboration, mainly used to observe the role of posts and interaction activities in supporting participants to learn, is to survey whether post contents may help others to solve problems herein. The learning collaboration of a_i is defined as follows:

$$\text{Lc}_I(a_i) = \frac{\text{HelpPostNum}(a_i)}{\text{TotalPostNum}(a_i)} \times \frac{\text{BeneficiaryNum}(a_i)}{N}, \quad (8)$$

where $\text{HelpPostNum}(a_i)$, $\text{TotalPostNum}(a_i)$, and $\text{BeneficiaryNum}(a_i)$, respectively, is the number of helpful posts of a_i , total number of posts, and the number of the beneficiaries from the posts of a_i .

The present difficulty is how to confirm whether a post of a_i may help others to solve problems. There is also similar research, including that on manual confirmation, which is

time and labour consuming and undesirable for massive corpus, and that on automatic confirmation, which identifies answer-question in forums in combination with rules and forum structure and achieves good results in extraction experiments [21]. In addition, this is also confirmed by vote in forums [22]. With the second method, the paper confirms the data about helpful posts and beneficiaries in line with rules and in a statistical manner.

3.3.2. Correlation with the Thread. Correlation with the thread title is the correlation of replies to the topic discussed in the main post. In the process of discussion in online learning forums, learners often post some irrelevant comments about question B in a post where question A is discussed. A superposter must not or rarely do so and should

$$Ec_I(a_i) = \frac{\text{KnowledgePointNum}(a_i, C_j)}{\text{TotalKnowledgePointNum}(C_j)} \times \frac{\text{KnowledgePostNum}(a_i, C_j)}{\text{TotalPostNum}(a_i, C_j)}, \quad (10)$$

where $\text{KnowledgePointNum}(a_i, C_j)$, $\text{KnowledgePostNum}(a_i, C_j)$, and $\text{TotalPostNum}(a_i, C_j)$, respectively, are the frequency of the knowledge points of C_j included in the posts posted by a_i , number of the posts which contain at least 1 knowledge point, and the total number of the posts posted in the C_j forum. $\text{TotalKnowledgePointNum}(C_j)$ is the total

comment in accordance with threads. Therefore, the correlation of a reply of a_i to topic is calculated as follows:

$$Ct_I(a_i) = \frac{\text{CorrNum}(a_i)}{\text{TotalReplyNum}(a_i)}, \quad (9)$$

where $\text{CorrNum}(a_i)$ and $\text{TotalReplyNum}(a_i)$, respectively, are the number of the replies of a_i considered relevant to the target topic (main post) and the total number of the replies of a_i . The correlation between a reply of a_i and the target thread is calculated with the cosine value.

3.3.3. Expertise of Content. Expertise of content is the course knowledge points involved in a post, by which the index of expertise of content in the posts of a_i in the C_j forum can be calculated as follows:

frequency of knowledge points of C_j in the forum. Each knowledge point which appears for 1 or 0 times is not counted repeatedly. $Ec_I(a_i)$ is the index of educational content of the posts sent by a_i in the C_j ($j = 1, \dots, K$) forum.

Accordingly, the content quality of a_i can be calculated as follows:

$$Cq_Index(a_i) = \mu_1 \times Lc_I(a_i) + \mu_2 \times Ct_I(a_i) + (1 - \mu_1 - \mu_2) \times Ec_I(a_i), \quad (11)$$

where μ_i , ($i = 1, 2$) are weighting parameters.

3.4. Superposter Index. With the MIN-MAX method, the paper normalizes the results of Le_Index , Cq_Index , and Sns_Index . For example, Le_Index can be normalized with the following formula to an extent that realizes the result within the range from 0 to 100:

$$NLe_Index(a_i) = \frac{\text{Le_Index}(a_i) - \text{MIN}(\text{Le_Index}(a_j)_{j=1}^N)}{\text{MAX}(\text{Le_Index}(a_j)_{j=1}^N) - \text{MIN}(\text{Le_Index}(a_j)_{j=1}^N)} \times 100. \quad (12)$$

Similarly, we may obtain NCq_Index and $NSns_Index$.

In conclusion, superposter index ($Super_I$) is calculated as follows:

$$Super_I = (1 - \alpha - \beta) \times NLe_Index + \alpha \times NCq_Index + \beta \times NSns_Index, \quad (13)$$

where α and β are weighting parameters which may be set according to the actual situation.

So Algorithm 1 can be described as ALG_Super_1 (see Algorithm 1, ALG_Super_1).

4. Experimental Result and Analysis

4.1. Data Set. The data in the paper are downloaded from the Q&A forum [23] for online learning course Computer Application Foundation. The dataset includes 7494 subject, 22369 posts, and 6747 participants (including 6712 learners

and 35 teachers). Among the 35 teachers, 28 were found to meet the defined conditions of a superposter, through sampling and analysis of the data about their posts (see Table 2). Therefore, 28 teachers were considered as superposters and identified with the method proposed in the paper; there were 7 teachers unqualified to be superposters for 4 teachers who posted 2 posts each and 3 teachers who posted 1 post each. In addition, to count the number of knowledge points in posts, the paper constructs an online unified examination knowledge point set based on the Fundamentals of Computer Application [24].

Input: the post dataset of online learners, D ; training parameter set, P .
Output: Superposter index, $Super_I$.

- (1) $A \leftarrow$ Counting online learners.
- (2) For each online learner a_i
 - (1) Computes $Sns_Index(a_i)$ by using formula (4);
 - (2) Computes $Le_Index(a_i)$ by using formula (7);
 - (3) Computes $Cq_Index(a_i)$ by formula (11);
 - (4) With the MIN-MAX method, normalizes the results of $Le_Index(a_i)$, $Cq_Index(a_i)$ and $Sns_Index(a_i)$:
 $NSns_Index(a_i) \leftarrow Sns_Index(a_i)$;
 $NCq_Index(a_i) \leftarrow Cq_Index(a_i)$;
 $NLe_Index(a_i) \leftarrow Le_Index(a_i)$;
- (5) Computes superposter index $Super_I(a_i)$ by using formula (13).
- (3) Rank $Super_I$;
- (4) Return $Super_I$.

ALGORITHM 1: ALG_Super_1: framework of computing superposter index.

4.2. Evaluation. There have been no mature and recognized methods of assessing superposter identification. In this section, evaluation is made with the following indexes [11].

The accuracy of TOP M ($P@M$) is as follows:

$$P@M = \frac{\text{the number of correct super posters in TOP } M}{M}. \quad (14)$$

The average accuracy of TOP M ($Avg P@M$) is as follows:

$$Avg P@M = \frac{\sum_{i=1}^M P@i}{M}. \quad (15)$$

4.3. Result Analysis. With the three feature indexes of the model including language expression (L), content quality (C), and network interaction structure (S), the paper makes a test on the effect of identifying the 28 superposters. Through repeated weighting tests on data, the weighting parameters are set in Table 3, and the results and statistical analysis are described in Tables 4 and 5, respectively. Comparison of the results of our algorithm with the PageRank algorithm (PR) is shown in Tables 5 and 6 and Figures 1–3.

According to Table 5, ① the model indexes achieve good results of superposters identification, and the model is very effective in application in the dataset. ② In terms of the identification result achieved by each feature, content quality, which realizes the average accuracy of over 0.9, is considered best. Language expression is just as the N-gram model, in which only 14 superposters are correctly identified among TOP 28, and is considered good. Social network structure, a common index in social network analysis by which 22 are correctly identified among TOP 28, is considered better. ③ Although a single feature is unable to perform well, their combinations may realize striking effects: with the combinations like LC, CS, and LCS, 24 superposters are correctly identified among TOP 28. With the combinations like LC and LCS, all of the superposters can be identified among TOP 15,

which undoubtedly proves that the feature designs are rational and effective. ④ The single feature L performs averagely, but its combination with other features performs well, especially LC. We are confused about whether this means that the two models are mutually complementary as part of contents in terms of structure. ⑤ Among TOP 28, 14 superposters are identified by L and LS models. This shows that language expression greatly depends on the length of text and is not sensitive to identification of the superposters of short text although consideration has been given to avoidance of this case in design. Since the number of the superposters identified by LS is less than that by S, we are confused about whether this means that L and S have something in common. However, there is a difference between L as a content-based result and S as a graph-based (network interaction) result in structure. We are confused that whether it is or not because the two models with different structures are mutually exclusive, which causes the result to deteriorate. Actually, this is also the case with the CS model, which achieves a result inferior to that C does, which we cannot explain in this study. Particularly, in the event of TOP 51, all of the 28 superposters can be identified. From Table 5, the trend chart for identification results and for average identification results achieved by each feature in different cases can be obtained (see Figures 1 and 2). Compared with the PageRank algorithm, the experiment result (see Tables 5 and 6 and Figures 1–3) of our algorithm (LCS) is better.

4.4. Discussion. In social networks, it is widely believed that out degree is as important as in degree; in the process of testing the weighting parameter of S, we found that when $\sigma = 1$, a good result was achieved locally and other values slightly improved; however, this was considered very unbalanced; that is, it only considered the replies to others' posts but neglected others' replies to own posts, leaving it not universal; through observation of data set, we found that there is a difference between the number of threads and the number of replies, especially in relation to the 28 teachers; in consideration of the generality of the model, through

TABLE 2: Comparison between two kinds of user of different identities in basic indexes.

Objects	Maximum number of posts sent	Maximum number of replies	Less than 10 posts	Average number of posts sent	Total number of Chinese characters in posts	Total number of words in posts
Teacher (35)	8617	8565	10	307.86	275340	147897
Learner (6712)	47	25	6658	1.73	548607	284833

TABLE 3: Parameters settings.

Parameters	Value
ϑ_1	0.2
ϑ_2, σ	0.5
μ_1	0.3
μ_2	0.4

TABLE 4: TOP 10 examples in L , C , and S .

User ID	L	C	S	Super_I
1527315	96.91504829	99.99880313	99.99430712	99.3807
1196662	98.79587171	53.61549801	0.868037451	46.82733
1191173	99.99825778	52.56897262	0.407785774	46.40647
1031382	97.07805875	49.47187606	0.023758924	44.15868
1191171	95.15300193	49.2480221	0.070274765	43.67569
1348	94.43927005	48.98495494	0.027054183	43.38845
511546	94.76087703	48.63108673	0.007664835	43.27002
1149	93.84371959	48.82328467	0.019282605	43.18617
1093505	92.56949365	48.49936007	0.020990493	42.76988
1237	93.38233965	48.13972648	0.002324243	42.74703

TABLE 5: Identification of superposters with each feature and their combinations.

Features	P@5	Avg P@5	P@10	Avg P@10	P@15	Avg P@15	P@20	Avg P@20	P@28	Avg P@28
L	0.8	0.96	0.7	0.80	0.6	0.74	0.55	0.69	0.5	0.65
C	1.0	1.0	0.9	0.94	0.93	0.93	0.9	0.93	0.86	0.92
S	0.8	0.91	0.9	0.89	0.87	0.88	0.9	0.88	0.79	0.86
LC	1.0	1.0	1.0	1.0	1.0	1.0	0.95	0.9975	0.86	0.98
LS	1.0	1.0	0.7	0.85	0.6	0.78	0.55	0.72	0.5	0.70
CS	0.8	0.96	0.9	0.92	0.93	0.92	0.9	0.92	0.86	0.91
LCS	1.0	1.0	1.0	1.0	1.0	1.0	0.9	0.98	0.86	0.95
PR	1	1	0.8	0.97	0.86	0.93	0.8	0.89	0.64	0.84

Note. $LC = 0.4L + 0.6C$, $LS = 0.3L + 0.7S$, $CS = 0.6C + 0.4S$, $LCS = 0.2L + 0.45C + 0.35S$.

TABLE 6: Recall and accuracy of identifying 28 superposters by LCS and PR.

	TOP 5	TOP 10	TOP 15	TOP 20	TOP 28	TOP 51	TOP 571
Recall-LCS	0.179	0.357	0.5367	0.64	0.86	1	—
Accuracy-LCS	1	1	1	0.9	0.86	0.549	—
Recall-PR	0.179	0.286	0.464	0.571	0.643	0.786	1
Accuracy-PR	1	0.8	0.86	0.8	0.643	0.43	0.049

comprehensive consideration, the paper sets $\sigma = 0.5$. Other parameters are set according to the optimum effects achieved by a single feature in experiments.

According to the intermediate results based on content quality, we found that some data were lost, such as the number of beneficiaries and helpful posts, especially as to learning supporters; for example, the posts sent or replies by teachers were related to the course and helpful to learners; therefore, all of the learners who participate in the

interaction were beneficiaries, and the posts of the teachers were considered helpful; however, the paper was unable to accurately obtain such information, leading to a loss of related data and affecting the identification of the teachers as superposters (the relationship between recall and accuracy can be seen in Table 6 and Figure 3). Nevertheless, through the model (LCS), all of the 28 superposters can be identified among TOP 51. It can be done by PageRank algorithm at top 571 cases.

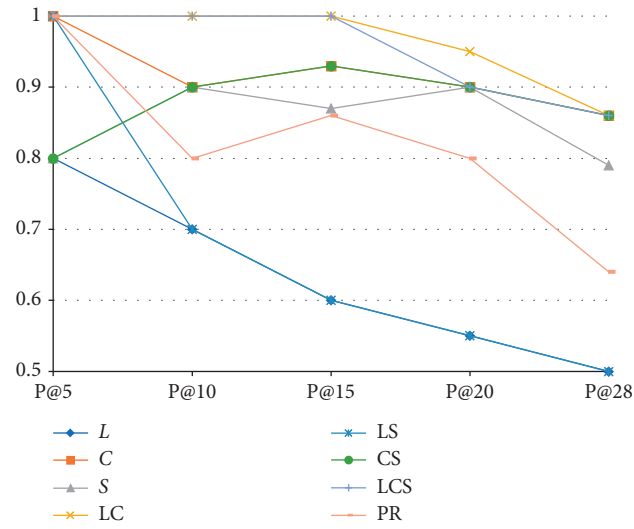


FIGURE 1: Trend chart for identification results achieved by each feature in different cases.

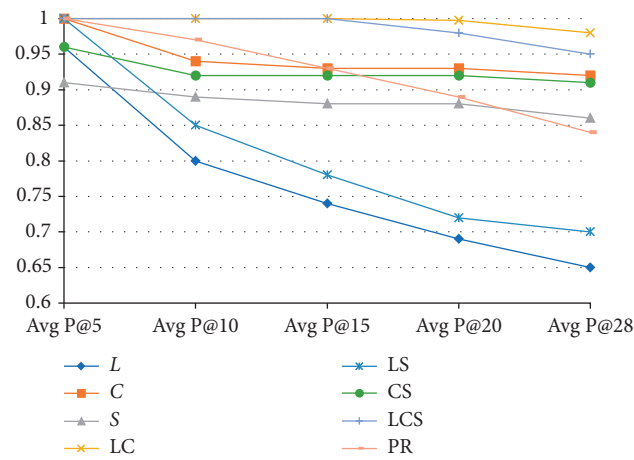


FIGURE 2: Trend chart for average identification results achieved by each feature in different cases.

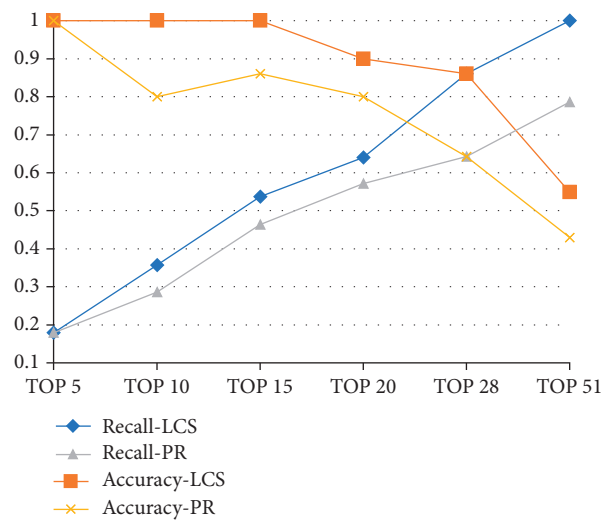


FIGURE 3: Trend chart of recall of and accuracy of identifying 28 superposters by LCS and PR.

5. Conclusion

Through analysis of the data on the posts sent by learners in online learning forums, the paper proposed a superposter identification model based on characters, words, and network interaction structure. First, through analysis of the network interaction of users based on graph structure, the paper calculated the out-degree centrality and in-degree centrality of each user node in networks, which involved both the interaction breadth and depth of each node, so as to determine its activity and importance in interactive networks. Then, learners' language expression was included in the identification framework, including word normalization, term normalization, and language elegance, by which the normalization of the words and terms used by learners and their basic ability to master language are judged. The third-dimensional feature is most important in online learning forums, i.e., content quality, which includes learning collaboration, correlation with the thread, and expertise of content. An online learning forum is designed to facilitate cooperation between learners and interaction in relation to learning contents. Learning collaboration mainly considers whether a post is helpful to others in study; correlation with the thread is to verify the correlation between a post and the topic discussed therein; expertise of content is to survey whether course knowledge points are included in a post. Accordingly, the three indexes work, respectively, in online learning forums on a targeted basis.

Although there are some deficiencies in the design of the superposter identification model, such as the need of repeated experiments on manual setting of weighting parameters in the process of calculation, a good result was achieved with the proposed method for identifying the preset 28 superposters. Considering that the method is easily realized and involves few calculations, and it is worthy to be applied in practical online learning systems.

Data Availability

Research data related to learners' personal privacy cannot be shared.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by China Scholarship Council and a general project of the National Natural Science Foundation of China (Event-Based Semantic Research on Educational Texts, Grant no. 61977032).

References

- [1] B. D. Wever, H. V. Keer, T. Schellens et al., "Roles as a structuring tool in online discussion groups: the differential impact of different roles on social knowledge construction," *Computers in Human Behavior*, vol. 26, no. 4, pp. 516–523, 2010.
- [2] S. Zha and C. Lee Ottendorfer, "Effects of peer-led online asynchronous discussion on undergraduate students' cognitive achievement," *American Journal of Distance Education*, vol. 26, no. 4, pp. 238–253, 2002.
- [3] N. Matsumura, Y. Ohsawa, and M. Ishizuka, "Identifying opinion leaders in the blogosphere," 2002.
- [4] X. Song, Y. Chi, K. Hino et al., "Mining and characterizing opinion leaders from threaded online discussions," 2007.
- [5] F. Li and T. C. Du, "Who is talking? An ontology-based opinion leader identification framework for word-of-mouth marketing in online social blogs," *Decision Support Systems*, vol. 51, no. 1, pp. 190–197, 2011.
- [6] Y. Cho, J. Hwang, and D. Lee, "Identification of effective opinion leaders in the diffusion of technological innovation: a social network approach," *Technological Forecasting and Social Change*, vol. 79, no. 1, pp. 97–106, 2012.
- [7] W. Zhang, H. He, and B. Cao, "Identifying and evaluating the internet opinion leader community based on k-clique clustering," *Neural Computing and Applications*, vol. 25, no. 3–4, pp. 595–602, 2014.
- [8] N. Ma and Y. Liu, "SuperedgeRank algorithm and its application in identifying opinion leader of online public opinion supernetwork," *Expert Systems with Applications*, vol. 41, no. 4, pp. 1357–1368, 2014.
- [9] P. Jia, A. Mirtabatabaei, N. E. Friedkin, and F. Bullo, "Opinion dynamics and the evolution of social power in influence networks," *Siam Review*, vol. 57, no. 3, pp. 367–397, 2015.
- [10] L. Yang, Y. Qiao, Z. Liu, J. Ma, and X. Li, "Identifying opinion leader nodes in online social networks with a new closeness evaluation algorithm," *Soft Computing*, vol. 22, no. 2, pp. 453–464, 2018.
- [11] X.-H. Fan, J. Zhao, F. A. N. G. Bin-Xing et al., "Influence diffusion probability model and utilizing it to identify network opinion leader," *Chinese Journal of Computers*, vol. 36, no. 2, pp. 360–367, 2013.
- [12] A. E. Reppel, I. Szmigin, and T. Gruber, "The iPod phenomenon: identifying a market leader's secrets through qualitative marketing research," *Journal of Product & Brand Management*, vol. 15, no. 4, pp. 239–249, 2006.
- [13] I. Himmelboim, E. Gleave, and M. Smith, "Discussion catalysts in online political discussions: content importers and conversation starters," *Journal of Computer-Mediated Communication*, vol. 14, no. 4, pp. 771–789, 2009.
- [14] Z. Zhai, H. Xu, and P. Jia, "Identifying opinion leaders in BBS," 2008.
- [15] D. K. Kim, A. C. James, and G. J. Shepherd, *Identifying Opinion Leaders by Using Social Network Analysis: A Synthesis of Opinion Leadership Data Collection Methods and Instruments*, The Faculty of the Scripps College of Communication of Ohio University, Columbus, OH, USA, 2007.
- [16] F. Bodendorf and C. Kaiser, *Detecting Opinion Leaders and Trends in Online Communities*, IEEE International Conference on Digital Society, New York, NY, USA, 2010.
- [17] J.-X. Cao, C. H. E. N. Gao-jun, W. U. Jiang-lin et al., "D multi-feature based opinion leader mining in social networks," *Acta Electronica Sinica*, vol. 44, no. 4, pp. 898–905, 2016.
- [18] Y. Li, S. Ma, Y. Zhang, R. Huang, and F. Kinshuk, "An improved mix framework for opinion leader identification in online learning communities," *Knowledge-Based Systems*, vol. 43, no. 2, pp. 43–51, 2013.
- [19] J. Huang, A. Dasgupta, A. Ghosh et al., "Superposter behavior in MOOC forums," 2014.
- [20] C. Romero, M.-I. López, J.-M. Luna, and S. Ventura, "Predicting students' final performance from participation in

- on-line discussion forums,” *Computers & Education*, vol. 68, pp. 458–472, C, 2013.
- [21] B. Wang, B. Liu, C. Sun et al., “Extracting Chinese question-answer pairs from online forums,” *Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics*, vol. 23, 2009.
- [22] P. Jurczyk and E. Agichtein, “Discovering authorities in question answer communities by using link analysis,” 2007.
- [23] <http://zjy.ccnue.edu.cn/jyfw/wljy.htm>.
- [24] National College Network Education Examination Committee Office, *Fundamentals of Computer Application*, Tsinghua University Press, London, UK, 2013.

Research Article

Forecast and Analysis of Coal Traffic in Daqin Railway Based on the SARIMA-Markov Model

Cheng Zhang ¹ and Shouchen Liu ^{1,2}

¹School of Transportation and Logistics, East China Jiao Tong University, Nanchang 330013, China

²School of Business Administration, Fujian Business University, Fuzhou 350012, China

Correspondence should be addressed to Shouchen Liu; 20426164@qq.com

Received 9 February 2020; Revised 2 December 2020; Accepted 11 December 2020; Published 23 December 2020

Academic Editor: Hongju Cheng

Copyright © 2020 Cheng Zhang and Shouchen Liu. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

With the continuous advancement of China's supply-side structural reform, the country's energy consumption structure has undergone considerable changes, including an overall reduction in fossil energy use and a rapid increase in clean energy application. In the context of China's coal overcapacity, port and rail capacities are difficult to change in the short term. This study forecasts the monthly coal traffic of Daqin Railway on the basis of the seasonal autoregressive integrated moving-average Markov model and then uses the monthly coal transport data of this railway from September 2009 to November 2019 as samples for model training and verification. Coal traffic from December 2019 to September 2020 is accurately predicted. This study also analyzes the effects of China's industrial structure adjustment, clean energy utilization, and low-carbon usage on the coal transport volume of Daqin Railway. In addition, the characteristics of seasonal fluctuation and the development trend of Daqin Railway's coal traffic are explored. This study provides a reference for adjusting the train operation chart of Daqin Railway's coal transport and developing a special coal train operation plan. It can determine the time of coal transport peak warning, improve the efficiency of coal transport management, and eventually realize a reasonable allocation of resources for Daqin Railway.

1. Introduction

Accounting for 1% of the length of China's railways, Daqin Railway comprises 20% of the country's national railway and 13% of its national coal transport volume. Daqin Railway holds and even constantly breaks the record for the fastest train running speed, the highest running density, the largest single railway volume, and the best transportation efficiency. It is a strategic artery of China's "west-to-east coal transport," which continuously carries coal to the Bohai Sea at a rate of 6.3 tons per second. However, under the background of China's coal production capacity, port and railway transport capacities are excessive and difficult to change within a short period. Moreover, the competition for China's coal supply transport channel is becoming increasingly fierce, and consequently, Daqin Railway is expected to experience pressure from new industry competition. Historical data can be used to predict the changing trend of coal traffic

in Daqin Railway. The railway transportation department can then adjust the train operation chart and formulate a coal train operation plan in accordance with the predicted coal traffic.

At present, local and foreign scholars have focused on railway freight forecasts. Many research achievements have been reported in the prediction of railway freight volume. Commonly used methods include the adaptive flocking algorithm [1], support vector machines [2], the gray model [3], time series models [4], neural networks [5], combination of models [6], multiple models [7], and regression analysis models [8]. Liu and Yu [9] used the seasonal autoregressive integrated moving-average (SARIMA) model to predict and analyze railway freight volume. Zhang [10] applied a time series model and a neural network to predict the annual freight volume of the Guangzhou-Shenzhen railway line. Zhao [11] solved an autoregressive integrated moving-average (ARIMA) model by using EViews software and predicted

China's railway freight volume. Huang et al. [12] and Yuan et al. [13] analyzed the considerable error produced by the gray Verhulst model in predicting railway freight volume and used a Markov chain model to modify the prediction results of the Verhulst model, improving prediction accuracy. Zhang and Zhou [14] used the gray forecast-Markov chain-qualitative analysis method to predict railway freight volume. Tang [4] constructed an improved gray MARKOV prediction model and predicted the future freight demand of China's railways. In Milenković et al. [15], the time series of the monthly passenger flow in Serbian railways from January 2004 to June 2014 was fitted and predicted using the SARIMA method. In Tang and Deng [16], an ARIMA model was established and R programming language was used to solve this model to make reasonable predictions of the future development trend of civil aviation passenger transport.

On the basis of existing studies, local and foreign scholars have conducted considerable research on railway freight transport prediction by using different methods and from various perspectives. However, relatively few studies on railway coal transport prediction have been performed in the context of China's coal production capacity, excessive railway transport capacity, and fierce competition among coal transport channels. In the current work, the actual coal transport volume situation of Daqin Railway is considered, and the SARIMA-Markov model is adopted to predict the monthly coal transport volume of Daqin Railway. This research analyzes the economic and market reasons for coal transportation in Daqin Railway and explores the seasonal fluctuation characteristics of coal transportation in this railway. The results of this study will provide an important reference for the managers of Daqin Railway to adjust their train operation chart and formulate a special operation plan for coal trains. The peak time of coal transport can be determined, and the efficiency of coal transport management can be improved.

The remainder of this paper is organized as follows. In Section 2, we provide the seasonal inspection data of monthly coal volume in Daqin Railway. In Section 3, we describe the SARIMA-Markov prediction model in detail. In Section 4, we verify the applicability of the SARIMA-Markov model by selecting the monthly coal volume of Daqin Railway from January 2009 to September 2019 as sample data for testing the model and forecasting future trends. Finally, we summarize the study in Section 5.

2. Seasonal Inspection of the Monthly Coal Traffic Volume of Daqin Railway

Daqin Railway provides a steady momentum for China's sustained economic development, and it has become a landmark among China's railways since the country's reform and opening up 40 years ago. In the monthly forecast of the coal traffic volume of Daqin Railway, the historical data of the railway's monthly coal traffic volume should be analyzed, and then an appropriate forecast model should be developed. In the current study, the monthly coal traffic volume of Daqin Railway from January 2009 to September 2019 is selected as the

observation data, and the time chart of the coal traffic volume of Daqin Railway is drawn using EVIEWS 10.0, as shown in Figure 1.

As shown in Figure 1, the monthly coal traffic of Daqin Railway exhibited a linear growth trend from 2009 to 2014. As a pilot industry of the supply-side structural reform, the State Council and the National Development and Reform Commission issued corresponding policies to cut coal capacity from 2015 to 2016, resulting in a linear decrease in the coal transport volume of Daqin Railway. After 2017, remarkable achievements were made in adjusting the industrial structure, and the overall benefit of the coal industry recovered steadily. The coal volume of Daqin Railway presented an increasing trend. In the short term, the monthly coal traffic of Daqin Railway exhibits evident seasonal fluctuation characteristics, with a cycle of 12 months. When constructing a prediction model, seasonal, trend growth, and random interference factors should be considered. In addition, the parameters of a matching prediction model should be selected to reduce prediction errors.

3. Modeling the SARIMA-Markov Prediction Model

3.1. SARIMA Prediction Model. The SARIMA model is derived from the ARIMA model. Its basic form is SARIMA $(p, d, q)(P, D, Q)^S$. The monthly coal transport volume time series of Daqin Railway is $\{x_t, t = 0, \pm 1, \pm 2, \dots\}$. The autoregressive (AR) model is as follows:

$$x_t = c + \varphi_1 x_{t-1} + \varphi_2 x_{t-2} + \dots + \varphi_p x_{t-p} + \varepsilon_t. \quad (1)$$

The moving-average (MA) model is as follows:

$$x_t = c + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \dots - \theta_q \varepsilon_{t-q}, \quad (2)$$

where c is a constant, ε_t is the residual sequence, and $\varepsilon_t \sim N(0, \sigma^2)$. When p is the lag order of the time series, the model is referred to as the AR(p) model. When q is the lag order of the residual sequence, the model is referred to as the MA(q) model. Equations (1) and (2) can be simplified as follows:

$$\begin{aligned} \varphi(B)x_t &= c + \varepsilon_t, \\ x_t &= c + \theta(B)\varepsilon_t, \end{aligned} \quad (3)$$

where B is the backward shift operator, and

$$\begin{aligned} B^j x_i &= x_{i-j}, \\ \varphi(B) &= 1 - \varphi_1 B - \varphi_2 B^2 - \dots - \varphi_p B^p, \\ \theta(B) &= 1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q. \end{aligned} \quad (4)$$

Models AR(p) and MA(q) are combined to produce ARMA(p, q),

$$\begin{aligned} x_t &= c + \varphi_1 x_{t-1} + \varphi_2 x_{t-2} + \dots + \varphi_p x_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} \\ &\quad - \theta_2 \varepsilon_{t-2} - \dots - \theta_q \varepsilon_{t-q}. \end{aligned} \quad (5)$$

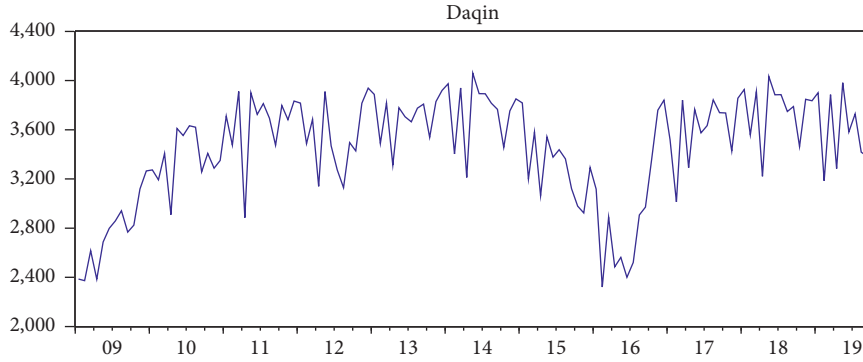


FIGURE 1: Time chart of the monthly coal traffic of Daqin Railway.

Equation (5) is simplified to the following:

$$\varphi(B)x_t = c + \theta(B)\varepsilon_t. \quad (6)$$

The time series should be stationary during analysis; otherwise, false regression will be produced, leading to unreliable predicted results. Obtaining a strictly stationary time series is difficult, and thus, a time series is required to be weakly stationary. The difference method is typically used to change a weakly stationary time series into a stationary sequence after one difference; however, other time series may require multiple differences. A stationary sequence with d -order difference is called a d -order difference sequence. A d -order difference sequence is applied to the ARMA(p, q) model to form the ARIMA(p, d, q) model, which is represented by the following:

$$\varphi(B)(1-B)^d x_t = c + \theta(B)\varepsilon_t. \quad (7)$$

Considering the periodicity of a time series, the SARIMA model can be obtained from the seasonal difference and the seasonal parameters of the ARIMA model. The general form of the SARIMA model is as follows:

$$\varphi(B)\Phi(B^S)(1-B)^d(1-B^S)^D x_t = c + \theta(B)\Theta(B^S)\varepsilon_t, \quad (8)$$

where s is the period of a time series, D is the order of seasonal difference, and B^S is the seasonal shift operator. Then, $\Phi(B^S) = 1 - \Phi_1 B - \Phi_2 B^2 - \dots - \Phi_p B^p$, $\Theta(B^S) = 1 - \Theta_1 B - \Theta_2 B^2 - \dots - \Theta_q B^q$.

The SARIMA model is denoted as SARIMA(p, d, q)(P, D, Q)^S, where d is the difference order of each period, D is the order of seasonal difference, p is the autoregressive order, q is the moving-average order, P is the seasonal autoregressive order, and Q is the seasonal moving-average order.

The steps of the SARIMA prediction model are as follows:

- (1) The sample data sequence is stabilized. Historical data for the observation period are selected as the sample sequence x_t . The first-order difference of x_t is determined to obtain $dx_t = D(x_t, d, S)$. The trend component in x_t is extracted, where d is the

difference order and S is the number of periodic difference steps.

- (2) The SARIMA(p, d, q)(P, D, Q)^S model is identified. The parameters (p, d, q)(P, D, Q) of the model are determined. The correlation of the difference sequence dx_t is analyzed, and the possible values of p and q are preliminarily identified in accordance with the truncated and trailing autocorrelation coefficient and partial correlation coefficient of dx_t [17]. Then, the significance of the model parameters is tested. The index values of R^2 , the Akaike information criterion (AIC), and the Schwarz criterion (SC) are compared. The optimal model parameters are identified.
- (3) Model adaptability is tested, and parameters are estimated. The correlation of the fitting residual sequence ε_t is analyzed to check if it is a white noise sequence. Whether the SARIMA(p, d, q)(P, D, Q)^S model fully extracts the useful information contained in sequence x_t is verified. If the correlation test result of ε_t is significant, then ε_t is not a white noise sequence, and the model cannot be adopted even if its evaluation index value is higher. By contrast, when the model passes the adaptability test, the least squares method is used to estimate model parameters.
- (4) The model's predictive power is evaluated. SARIMA(p, d, q)(P, D, Q)^S is evaluated. The evaluation index of the predictive power of the model in the sample period is $\text{MAPE} = (1/n)|(\hat{x}_t - x_t)/x_t| \times 100$. If the prediction effect of $\text{MAPE} \leq 5\%$ is excellent, then the prediction effect of $5\% \leq \text{MAPE} \leq 10\%$ is good, that of $10\% \leq \text{MAPE} \leq 20\%$ is qualified, and that of $10\% \leq \text{MAPE} \leq 20\%$ is unqualified [18].

3.2. Markov Prediction Model. The Markov model exerts a nonposterior effect. For a random time series $\{X(t), t \in T\}$, where T is a discrete time set, the nonposterior effect is expressed as follows [19]:

$$P\{X_{n+1} = i_{n+1} | X_1 = i_1, X_2 = i_2, \dots, X_n = i_n\} = P\{X_{n+1} = i_{n+1} | X_n = i_n\}. \quad (9)$$

The steps of the Markov prediction model are as follows:

- (1) The residual sequence e_i is solved as follows:

$$e_i = x_i - f_i, \quad i = 0, 1, 2, \dots, \quad (10)$$

where x_i denotes the original data, and f_i is the predicted value of the SARIMA model [20].

- (2) The states are divided. e_i is divided into r states, with the same spacing width from large to small. The upper limit of the j state in step i is as follows:

$$U_{ij} = \min e_i + \frac{j-1}{r} (\max e_i - \min e_i), \quad (11)$$

and lower limit is as follows:

$$L_{ij} = \min e_i + \frac{j}{r} (\max e_i - \min e_i). \quad (12)$$

- (3) The probability of state transition in k steps is solved. Suppose the state space is $\Omega = \{i_1, i_2, \dots, i_r\}$, and the probability of state transition in step k is $\{i_i = i_1, i_i = i_2, \dots, i_i = i_r\}$ [21].
- (4) The state transition matrix is built. A one-step transfer matrix that can reflect the probability transfer of various states between systems is denoted as follows:

$$P_{(0)} = \begin{pmatrix} p_{11} & \cdots & p_{1r} \\ \vdots & & \vdots \\ p_{r1} & \cdots & p_{rr} \end{pmatrix}. \quad (13)$$

- (5) The predicted value of the Markov model is expressed as follows:

$$\hat{x}_{t+1} = f_{t+1} - \sum_{i=1}^r a_i(t) z_i, \quad (14)$$

where \hat{x}_{t+1} is the predicted value of the SARIMA-Markov model at time $t+1$ [22].

$a_i(t)$ ($i = 1, 2, \dots, r$) is the probability for a one-step state transition of the row vector of the state transition matrix [22–24]. z_i ($i = 1, 2, \dots, r$) is the possible predicted value of each state interval. $z_i = (U_{ij} + L_{ij})/2$ is selected in this study [25].

4. Case Analysis

To verify the applicability of the SARIMA-Markov model, this study selected the monthly coal traffic volume of Daqin Railway from January 2009 to September 2019 as sample data to establish the prediction model. Evaluation indicators were adopted to evaluate the model. Among these, sample

data from October to November 2019 were used for the model test, and trend extrapolation was performed for the period of December 2019 to September 2020.

4.1. SARIMA Model

- (1) The data sequence x_t of the sample period is stabilized. EViews 10.0 software is used to find the first difference of x_t . The mean value of $dx_t = D(x_t, 1, 0)$ is approximately zero, and the trend characteristic disappears, as shown in Figure 2. The linear trend of sequence x_t is fully extracted by the first-order difference [26].

The autocorrelation diagram of the first-order difference sequence dx_t is provided in Figure 3. The autocorrelation coefficients near the 12th and 24th orders are significantly not 0, and thus, a seasonal feature with a period of 12 exists in the first-order difference sequence dx_t . This finding is consistent with the intuitive analysis results of the sequence diagram [27].

After the difference operation with a period of 12 on dx_t , a new sequence $d'x_t = D(x_t, 1, 12)$ is obtained [28]. As shown in Figure 4, the 12th-order autocorrelation coefficient of the second-order difference sequence $d'x_t$ is close to 0, indicating that the periodic factor dx_t is fully extracted [29–31].

An augmented Dickey-Fuller (ADF) test was performed on the second-order difference sequence $d'x_t$, and the results are provided in Table 1.

The unit root statistic is $ADF = -7.635505$, which is less than the critical value with a significance level of 1%. Thus, the null hypothesis that states that the second-order difference sequence $d'x_t$ has a unit root was rejected, indicating that the sequence $d'x_t$ is a stationary sequence. Accordingly, the parameters of the SARIMA $(p, d, q)(P, D, Q)^S$ model are $d = 1, D = 1, S = 12$.

- (2) The SARIMA $(p, d, q)(P, D, Q)^S$ model is determined [32]. As shown in Figure 4, x_t exhibits no significant correlation between different points in the same period. Thus, a simple seasonal model is established for the data sequence SARIMA $(p, 1, q)(1, 1, 1)^{12}$ of the sample period. The 1st-order autocorrelation coefficient of the differential stationary $d'x_t$ is significant, and its 2nd-order partial autocorrelation coefficient is also significant. The autocorrelation and partial autocorrelation diagrams are tailed; thus, five models can be built as follows:

$$\begin{aligned} & \text{SARIMA}(2, 1, 1)(1, 1, 1)^{12}, \\ & \text{SARIMA}(4, 1, 0)(1, 1, 1)^{12}, \\ & \text{SARIMA}(1, 1, 3)(1, 1, 1)^{12}, \\ & \text{SARIMA}(1, 1, 1)(1, 1, 1)^{12}, \\ & \text{SARIMA}(1, 1, 2)(1, 1, 1)^{12}. \end{aligned} \quad (15)$$

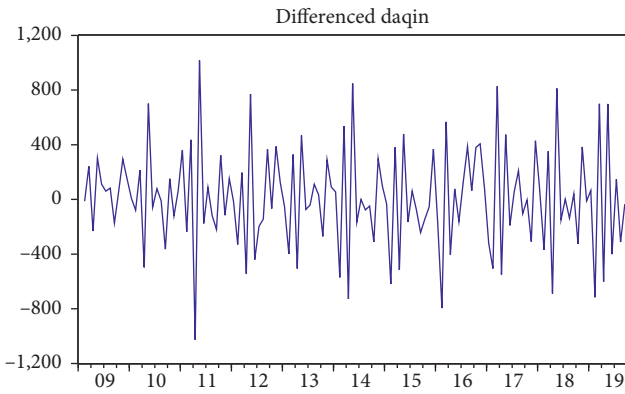


FIGURE 2: First-order difference sequence of x_t .

Sample: 2009M01 2019M09
Included observations: 128

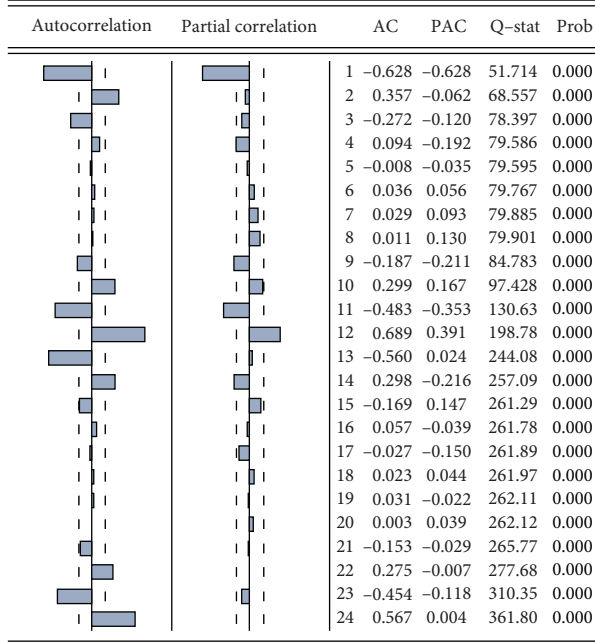


FIGURE 3: Autocorrelation of sequence dx_t .

The five models are tested for parameter significance, and the results are presented in Table 2.

R^2 is the goodness of fit of the entire model; the higher the R^2 value, the better the fit degree, where $R^2 \in [0, 1]$. AIC and SC are information criteria; the lower the value, the better the fitting degree of the model [33]. The test results of five compared models are provided in Table 2. The sequence model of data x_t in the sample period is SARIMA (4, 1, 2) (1, 1, 1)¹².

- (3) Model adaptability is tested, and parameters are estimated. In EViews 10.0, the SARIMA (4, 1, 2) (1, 1, 1)¹² model is used to obtain the fitting residual sequence ε_t , as shown in Figure 5.

Sample: 2009M01 2019M09
Included observations: 114

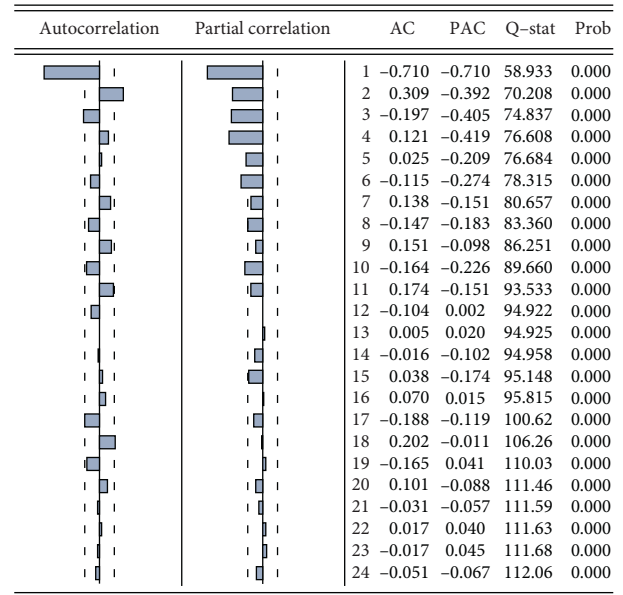


FIGURE 4: Autocorrelation of sequence $d'x_t$.

TABLE 1: ADF test results of sequence $d'x_t$.

	t-statistic	Prob. *
Augmented dickey-fuller test statistic	-7.635505	0.0000
Test critical values	1% level	-3.487550
	5% level	-2.886509
	10% level	-2.580163

TABLE 2: Significance test of model parameters.

Model	R^2	AIC	SC
SARIMA (2, 1, 1) (1, 1, 1) ¹²	0.3977	14.2007	14.2676
SARIMA (1, 1, 2) (1, 1, 1) ¹²	0.4073	14.1851	14.2519
SARIMA (1, 1, 3) (1, 1, 1) ¹²	0.4064	14.1866	14.2534
SARIMA (4, 1, 0) (1, 1, 1) ¹²	0.0088	14.6795	14.7241
SARIMA (4, 1, 2) (1, 1, 1) ¹²	0.6762	13.7281	13.8395

A white noise test is performed on the model residual ε_t of SARIMA (4, 1, 2) (1, 1, 1)¹²; that is, a random test of fitting residual sequence ε_t . The test results are presented in Figure 6.

As shown in Figure 6, the statistical value of Q is 8.963, and the associated probability is 0.34 > 0.05 from line $K = 12$. The null hypothesis states that no correlation exists and ε_t is accepted, indicating that ε_t is a white noise sequence. Therefore, the SARIMA (4, 1, 2) (1, 1, 1)¹² model passes the adaptability test, and its expression is as follows:

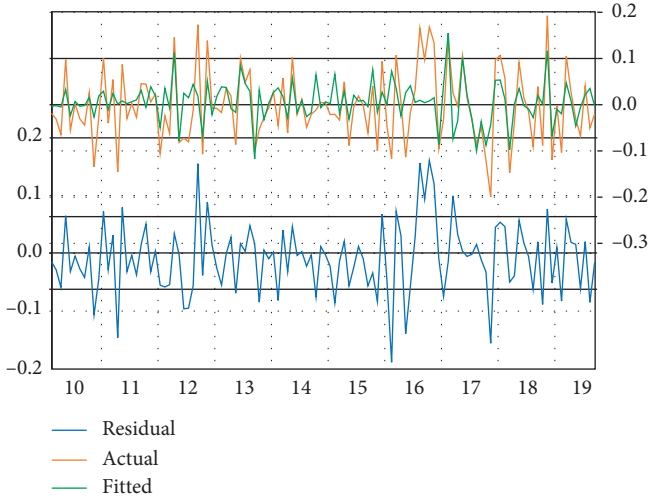


FIGURE 5: Fitting of the SARIMA(4, 1, 2)(1, 1, 1)¹² model.

Date: 11/20/19 Time: 13:22
 Sample: 2009M01 2019M09
 Included observations: 116
 Q-statistic probabilities adjusted for 4 ARMA terms

Autocorrelation	Partial correlation	AC	PAC	Q-stat	Prob
█	█	1 -0.1...	-0.1...	4.148...	
█	█	2 -0.0...	-0.0...	4.158...	
█	█	3 -0.1...	-0.1...	6.182...	
█	█	4 -0.0...	-0.0...	6.182...	
█	█	5 0.02...	-0.0...	6.236...	0.01...
█	█	6 -0.0...	-0.1...	7.317...	0.02...
█	█	7 0.07...	0.02...	7.989...	0.04...
█	█	8 -0.0...	-0.0...	8.049...	0.08...
█	█	9 0.02...	-0.0...	8.153...	0.14...
█	█	10 -0.0...	-0.0...	8.413...	0.20...
█	█	11 0.06...	0.05...	8.958...	0.25...
█	█	12 -0.0...	0.00...	8.963...	0.34...

FIGURE 6: Correlation test of the residual sequence ε_t in the SARIMA(4, 1, 2)(1, 1, 1)¹² model.

$$\begin{aligned}
 & \cdot (1 + 0.6569B + 0.9473B^2 + 0.2925B^3 - 0.0381B^4) \\
 & \cdot (1 - 0.148B^{12})\Delta_{12}\lg x_t \\
 & = (1 - 0.4657B - 0.8686B^2)(1 + 0.87B^{12})\varepsilon_t.
 \end{aligned}
 \tag{16}$$

(4) Model predictive power is assessed. In EViews 10.0, the SARIMA(4, 1, 2)(1, 1, 1)¹² model is used to predict the sample data of the coal traffic volume of Daqin Railway from January 2010 to September 2019. The predicted results are presented in Figure 7. An analysis model prediction level evaluation index, namely, mean absolute percentage error (MAPE) =

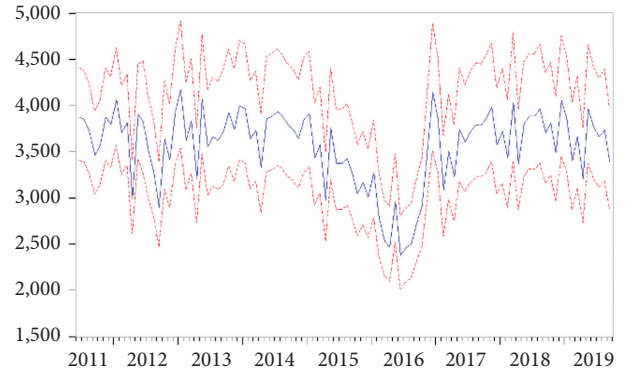


FIGURE 7: Prediction result of the SARIMA(4, 1, 2)(1, 1, 1)¹² model in the sample period.

5.1439%, is used to determine whether the model's predictive capability is good.

The SARIMA(4, 1, 2)(1, 1, 1)¹² model is used to predict the monthly coal transport volume of Daqin Railway from January 2009 to September 2019, and the results are provided in Table 3. The analysis of Table 3 indicates that the error of the single prediction model is still relatively large. To avoid this situation, the Markov model is adopted to correct the residual. This model can reduce the relative residual value and find the internal regularity.

4.2. SARIMA-Markov Model. In accordance with the steps of the Markov model, the range of the residual sequence $e_i \in [-6.884, 5.834]$ is first determined by dividing the residual sequence into E_1, E_2, E_3, E_4 state intervals. The boundary value of state interval E_1 is as follows:

$$U_{ij} = -6.884 + \frac{1-1}{4} (5.834 - (-6.884)) = -6.884,
 \tag{17}$$

$$L_{ij} = -6.884 + \frac{1}{4} (5.834 - (-6.884)) = -3.705.$$

Similarly,

$$\begin{aligned}
 E_2 & \in [-3.705, -0.525], \\
 E_3 & \in [-0.525, 2.655], \\
 E_4 & \in [2.655, 5.834].
 \end{aligned}
 \tag{18}$$

In accordance with the state distribution of the residual series, the frequency statistics of the one-step transfer from E_i to E_j are provided in Table 4.

The state transition probability matrix is as follows:

$$P_{(0)} = \begin{pmatrix} 0 & 0 & 0.6000 & 0.4000 \\ 0.0263 & 0.2895 & 0.4737 & 0.2105 \\ 0.0545 & 0.4000 & 0.4909 & 0.0545 \\ 0.0769 & 0.3846 & 0.5385 & 0 \end{pmatrix}.
 \tag{19}$$

We obtain the following:

TABLE 3: Statistical table of the SARIMA model's predicted value and residual sequence (in million tons).

Time	ACTUAL	SARIMA F	Error
2010/6	35.5380	36.3685	-0.8305
2010/7	36.3230	36.6030	-0.2800
2010/8	36.2110	36.1051	0.1059
2010/9	32.5760	36.0014	-3.4254
2010/10	34.0840	32.7257	1.3583
2010/11	32.8800	35.9553	-3.0753
2010/12	33.4990	34.0188	-0.5198
2011/1	37.1090	32.5815	4.5275
2011/2	34.7510	36.8717	-2.1207
2011/3	39.1160	36.4210	2.6950
2011/4	28.8430	35.7270	-6.8840
2011/5	39.0050	34.0413	4.9637
2011/6	37.2460	38.0273	-0.7813
2011/7	38.1220	38.9142	-0.7922
2011/8	36.9390	36.7624	0.1766
2011/9	34.7450	35.8810	-1.1360
2011/10	37.9690	35.2646	2.7044
2011/11	36.8160	38.5455	-1.7295
2011/12	38.3310	37.8582	0.4728
2012/1	38.1740	39.5214	-1.3474
2012/2	34.8760	37.6725	-2.7965
2012/3	36.8430	37.0296	-0.1866
2012/4	31.4010	31.2652	0.1358
2012/5	39.1010	37.1483	1.9527
2012/6	34.6950	38.1825	-3.4875
2012/7	32.7310	35.4198	-2.6888
2012/8	31.2900	32.1657	-0.8757
2012/9	34.9530	29.1189	5.8341
2012/10	34.2850	35.6686	-1.3836
2012/11	38.1570	34.3911	3.7659
2012/12	39.3840	38.9422	0.4418
2013/1	38.8740	41.9307	-3.0567
2013/2	34.8820	36.7719	-1.8899
2013/3	38.1690	37.8120	0.3570
2013/4	33.1170	32.6547	0.4623
2013/5	37.7950	39.3172	-1.5222
2013/6	37.0570	35.7993	1.2577
2013/7	36.6493	37.0051	-0.3558
2013/8	37.7550	36.3293	1.4257
2013/9	38.0900	36.7567	1.3333
2013/10	35.3830	39.4495	-4.0665
2013/11	38.2820	36.5611	1.7209
2013/12	39.1930	39.4069	-0.2139
2014/1	39.7410	40.0323	-0.2913
2014/2	34.0330	36.6904	-2.6574
2014/3	39.3990	37.0087	2.3903
2014/4	32.1220	33.8711	-1.7491
2014/5	40.6070	38.3192	2.2878
2014/6	38.9360	38.3357	0.6003
2014/7	38.9250	39.6389	-0.7139
2014/8	38.1540	38.4263	-0.2723
2014/9	37.6650	38.3145	-0.6495
2014/10	34.5620	37.6351	-3.0731
2014/11	37.5350	36.0587	1.4763
2014/12	38.5080	38.1776	0.3304
2015/1	38.1860	39.0928	-0.9068

TABLE 4: Statistics of the one-step state transfer frequency.

	$y_i^1 \rightarrow j$				$\sum y_i^1 \rightarrow j$
	$j=1$	$j=2$	$j=3$	$j=4$	
1	0	0	3	2	5
2	1	11	18	8	38
3	3	22	27	3	55
4	1	5	7	0	13

$$\begin{aligned}
 z_1 &= -5.2945, \\
 z_2 &= -2.115, \\
 z_3 &= 1.065, \\
 z_4 &= 4.245,
 \end{aligned}
 \tag{20}$$

from the formula $z_i = (U_{ij} + L_{ij})/2$. The prediction value of the SARIMA-Markov model is calculated in accordance with equations (13) and (14), and the prediction capability of the model is determined to be excellent by the analysis model's prediction level evaluation indicator, i.e., MAPE = 3.8009%. The fitting curves of the actual (ACTUAL), the predicted value of the SARIMA model (SARIMA F), and the predicted value of the SARIMA-Markov model (SARIMA-Markov F) are presented in Figure 8. As shown in the figure, the fitting effect of the SARIMA-Markov model is better than that of the SARIMA model (in million tons).

4.3. *Model Validation.* To verify the prediction accuracy of the established model, the SARIMA-Markov model was used to predict the monthly coal traffic volume of Daqin Railway in October and November 2019. The predicted values for October and November 2019 were 32.5864 and 38.7377 million tons, respectively, when the SARIMA(4, 1, 2)(1, 1, 1)¹² model was used. From Table 3, the state probability transfer vector in September 2019 is E_3 , and the initial state transfer vector is $x_0 = [0.0545 \ 0.4 \ 0.4909 \ 0.0545]$.

In accordance with the initial state probability transfer vector 1 and the state probability transfer matrix 2, the state probability transfer vectors of October and November 2019 can be calculated as follows:

$$\begin{aligned}
 x_1 &= x_0 * p_{(0)} = [0.0415 \ 0.3319 \ 0.4925 \ 0.1328], \\
 x_2 &= x_1 * p_{(0)} = [0.0458 \ 0.3432 \ 0.4954 \ 0.1133].
 \end{aligned}
 \tag{21}$$

Then, from equations (13) and (14), the predicted coal transport volume in October 2019 is 34.1619 million tons and the predicted coal transport volume in November 2019 is 32.8694 million tons. The results are provided in Table 5.

In the current study, mean absolute error (MAE) and MAPE were used to evaluate the fitting effect of the model, as indicated in Table 6. Both indexes were smaller for the optimized model than those before optimization. The fitting

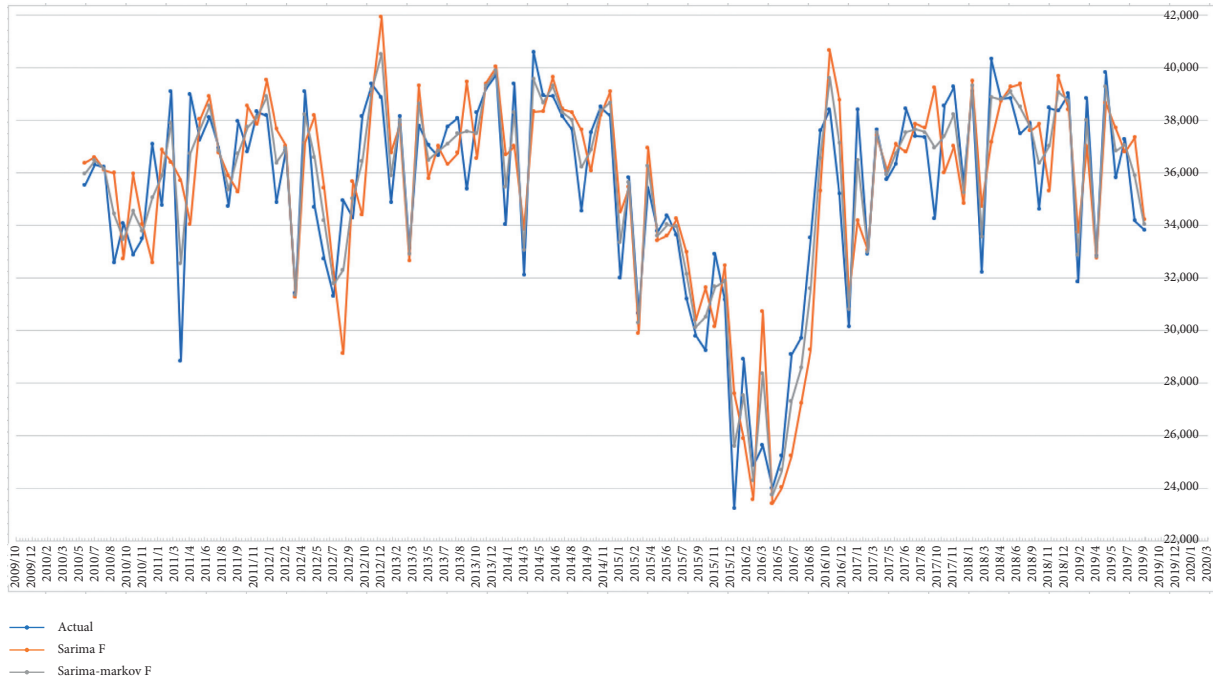


FIGURE 8: SARIMA-Markov model fitting diagram.

TABLE 5: Statistics of the prediction results (in million tons).

Time	Actual	SARIMA F	SARIMA-markov F
201910	36.4439	32.8313	34.1619
201911	35.4127	38.6018	32.8694

TABLE 6: Model test list (in million tons).

Forecast model	Model fitting		Model forecasting	
	MAE	MAPE	MAE	MAPE (%)
SARIMA	5.7611	5.1439	3.4003	9.4576
SARIMA-markov	4.2570	3.8009	2.4094	6.7126

TABLE 7: Statistics of the forecast results (in million tons).

Time	Company F	SARIMA F	SARIMA MARKOV F
2019/12 (E)	39.4429	40.5264	38.7993
2020/01 (E)	41.5839	39.4894	40.5889
2020/02 (E)	34.7971	36.6173	35.5534
2020/03 (E)	41.0039	40.6742	40.7300
2020/04 (E)	34.2827	34.2780	34.3346
2020/05 (E)	43.4109	41.4161	41.4787
2020/06 (E)	38.8871	40.6402	39.5333
2020/07 (E)	40.3224	39.7717	40.7116
2020/08 (E)	34.8113	39.5051	38.7080
2020/09 (E)	34.5971	34.6678	34.4061

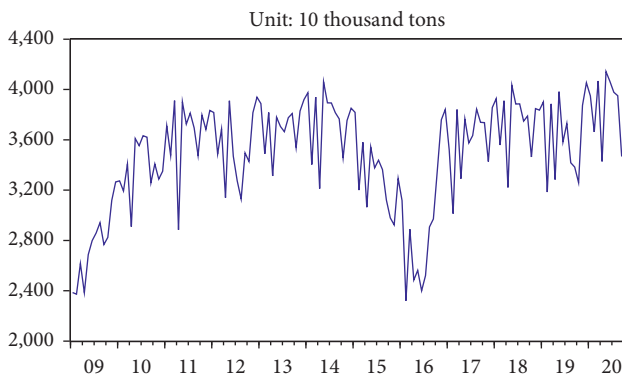


FIGURE 9: Prediction result of the SARIMA (4, 1, 2)(1, 1, 1)¹² model outside the sample period.

accuracy of the SARIMA-Markov model reached 96.1991%, which was higher than the prediction accuracy of the SARIMA model (94.8561%). The prediction accuracy of the

SARIMA-Markov model reached 93.2874%, which was higher than the prediction accuracy of the SARIMA model (90.5424%). Thus, the prediction accuracy of the SARIMA-Markov model is high and meets the requirements.

4.4. Model Prediction. The model was built on the basis of the monthly data of the coal traffic volume of Daqin Railway from January 2009 to September 2019. The predicted sequence x_f and the actual sequence x were drawn in the same diagram for comparison, as shown in Figure 9. In accordance with the state probability transition matrix P and the initial state transition vector x_0 , the state transition vector from December 2019 to September 2020 was determined. Simultaneously, the predicted values of SARIMA (4, 1, 2)(1, 1, 1)¹² from December 2019 to September 2020 were obtained.

The coal traffic volume of Daqin Railway from December 2019 to September 2020 was predicted using the SARIMA-Markov model. The results are provided in Table 7.

5. Conclusion

The SARIMA-Markov model was applied to the monthly coal traffic forecast of Daqin Railway. The SARIMA model comprehensively considered the influence of the seasonal correlation of coal traffic volume on Daqin Railway. Meanwhile, the Markov model used a residual sequence, state partition, and state transition matrix to modify the influence of the sample data on the predicted value. Compared with the simple use of the SARIMA model alone, the combination of the two models can ensure higher prediction accuracy, verifying the scientificity and feasibility of the model and providing a new method for the coal volume prediction of Daqin Railway. Moreover, the coal traffic volume of Daqin Railway from December 2019 to September 2020 was predicted via trend extrapolation, and the forecast results were analyzed as follows.

- (1) China's economy is less dependent on coal. As China's economy moves toward high-quality development, the growth rate of the country's gross domestic product will remain at approximately 6% in the next few years. Meanwhile, the Daqin Railway, as a strategic artery of China's "coal transport from the west to the east," will experience a growth rate of approximately 5%. This conclusion indicates that China's extensive economic development model based on coal energy consumption is beginning to weaken, and the economic form is developing toward the low-carbon, efficient, and green direction. China's coal dependence is decreasing due to the country's effort to adjust its industrial structure, encourage innovations, and increase the intensity of scientific and technological research and development.
- (2) Initial results have been achieved in optimizing the energy structure. The growth rate of the coal transport volume of Daqin Railway will gradually slow down under the influences of energy structure adjustment, industrial structure adjustment, non-fossil energy development, and other factors. This scenario shows a steady decline in the proportion of coal in energy consumption. The rapid development of low-cost nuclear, photovoltaic, and wind power will further reduce the demand for coal, which will account for a smaller share of the country's total energy consumption, while clean, renewable energy will obtain a larger share.
- (3) The process of providing heat from clean energy is proceeding in an orderly manner. China is a developing industrial country that uses coal boilers to provide heating during winter, particularly in the north of Qinling and Huai River. Under the background of the energy revolution, China has effectively promoted clean heating in the north and has replaced coal with gas and electricity for providing heat in an orderly manner, reducing the consumption of coal. In accordance with the local conditions, we expand the variety of clean heating methods to

ensure the balanced development of clean heating. In our future research, the SARIMA model will be combined with the random forest model to improve the accuracy of the model prediction.

Data Availability

Daqin Railway belongs to China Daqin Railway Co., Ltd., which is a listed company. The data required in this paper can be inquired in the monthly or annual financial statements of China Daqin Railway Co., Ltd., which belongs to the public information of the company.

Conflicts of Interest

None of the authors have any conflicts of interest.

References

- [1] Y. Yang, N. Xiong, N. Y. Chong, and X. Défago, "A decentralized and adaptive flocking algorithm for autonomous mobile robots," in *Proceedings of the 3rd International Conference on Grid and Pervasive Computing*, Kunming, China, May 2008.
- [2] X. F. Yu, *Research on the application of machine learning theory in railway freight volume prediction*, Ph.D. thesis, Beijing Jiaotong University, Beijing, China, 2016.
- [3] R. Li, M. R. Dai, and F. Z. Li, "Study on selection methods of key influence factors of railway freight transport volume based on grey correlation," *Railway Freight Transport*, vol. 33, no. 11, pp. 11–14, 2015.
- [4] Z. P. Tang, J. W. Zhu, and J. P. Sun, "Railway freight demand forecasting based on the improved grey MARKOV mod," *Railway Purchasing and Logistics*, vol. 10, no. 3, pp. 57–59, 2015.
- [5] D. Wang and G. J. Mi, "Prediction study of rail-way freight volume based on grey relational analysis and BP neural network," *Journal of Jiangnan University (Natural Science Edition)*, vol. 14, no. 1, pp. 80–84, 2015.
- [6] W. T. Zhu, "Forecasting railway freight volume based on improved BP neural network model," *Journal of Shijiazhuang Tiedao University (Natural Science Edition)*, vol. 27, no. 2, pp. 79–82, 2014.
- [7] Q. Zhang, C. Zhou, N. Xiong, Y. Qin, X. Li, and S. Huang, "Multimodel based incident prediction and risk assessment in dynamic cybersecurity protection for industrial control systems," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 46, no. 10, pp. 1429–1444, 2015.
- [8] Y. Yan and Z. K. Wu, "A study on a freight volume forecast model at harbin railway hub based on grey-linear regression model," *Railway Freight Transport*, vol. 36, no. 11, pp. 1–5, 2018.
- [9] M. T. Liu and J. L. Yu, "Forecast analysis of China railway freight volume based on SARIMA model," *Journal of Guizhou Normal College*, vol. 31, no. 12, pp. 43–47, 2015.
- [10] X. Zhang, *Research of railway freight volume forecasting based on data mining*, Ph.D. thesis, Southwest Jiaotong University, Chengdu, China, 2016.
- [11] N. Q. Zhao, "Forecast of railway freight volume based on ARIMA model," *Think Tank Era*, vol. 22, p. 187, 2019.
- [12] K. Huang, Q. Zhang, C. Zhou, N. Xiong, and Y. Qin, "An efficient intrusion detection approach for visual sensor networks based on traffic pattern learning," *IEEE Transactions on*

- Systems, Man, and Cybernetics: Systems*, vol. 47, no. 10, pp. 2704–2713, 2017.
- [13] S. Q. Yuan, B. Xue-ying, and W. Qi-cai, “Research on railway freight volume prediction based on Markov verhulst model,” *Railway Standard Design*, vol. 60, no. 10, pp. 27–30, 2016.
- [14] C. Zhang and X.-F. Zhou, “Prediction of railway freight volumes based on gray forecast-Markov chain-qualitative analysis,” *Journal of the China Railway Society*, vol. 29, no. 5, pp. 15–21, 2007.
- [15] M. Milenković, L. Švadlenka, and V. Melichar, “SARIMA modelling approach for railway passenger flow forecasting,” *Transport*, vol. 33, no. 5, pp. 1113–1120, 2018.
- [16] X. Tang and G. Deng, “Prediction of civil aviation passenger transportation based on ARIMA model,” *Open Journal of Statistics*, vol. 6, no. 5, pp. 824–834, 2016.
- [17] A. Shahzad, M. Lee, Y.-K. Lee et al., “Real time MODBUS transmissions and cryptography security designs and enhancements of protocol sensitive information,” *Symmetry*, vol. 7, no. 3, pp. 1176–1210, 2015.
- [18] W. Wu, N. Xiong, and C. Wu, “Improved clustering algorithm based on energy consumption in wireless sensor networks,” *IET Networks*, vol. 6, no. 3, pp. 47–53, 2017.
- [19] W. Zhang, D. Chen, H. Si, and N. N. Xiong, “RTDCM: a coding preemption collection system for key data prioritization with hierarchical probability exchange mechanism in mobile computing,” *IEEE Access*, vol. 8, pp. 4629–4639, 2020.
- [20] C. Chen, N. N. Xiong, X. Guo, and J. Ren, “The system identification and prediction of the social earthquakes burst in human society,” *IEEE Access*, vol. 8, pp. 103848–103859, 2020.
- [21] M. Wu, L. Zhong, L. Tan, and N. Xiong, “The sequential fusion estimation algorithms based on gauss-Newton method over multi-agent networked systems,” *IEEE Access*, vol. 8, pp. 114315–114329, 2020.
- [22] B. Yin, X. Wei, J. Wang, N. Xiong, and K. Gu, “An industrial dynamic skyline based similarity joins for multidimensional big data applications,” *IEEE Transactions on Industrial Informatics*, vol. 16, no. 4, pp. 2520–2532, 2020.
- [23] W. Guo, Y. Shi, S. Wang, and N. N. Xiong, “An unsupervised embedding learning feature representation scheme for network big data analysis,” *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 1, pp. 115–126, 2020.
- [24] Y. Zhang, J. Chen, B. Liu et al., “COVID-19 public opinion and emotion monitoring system based on time series thermal new word mining,” *Computers, Materials and Continua*, vol. 64, no. 3, p. 11458, 2020.
- [25] A. Fu, X. Zhang, N. Xiong, Y. Gao, and H. Wang, “VFL: a verifiable federated learning with privacy-preserving for big data in industrial IoT,” *IEEE Transactions on Industrial Informatics*, p. 13585. In press.
- [26] G. Liu, C. Guo, L. Xie, W. Liu, N. Xiong, and G. Chen, “An intelligent CNN-VAE text representation technology based on text semantics for comprehensive big data,” *CoRR Abs/2008*, In press.
- [27] Y. Ren, W. Liu, T. Wang, X. Li, N. N. Xiong, and A. Liu, “A collaboration platform for effective task and data reporter selection in crowdsourcing network,” *IEEE Access*, vol. 7, pp. 19238–19257, 2019.
- [28] H. Cheng, Z. Xie, Y. Shi, and N. Xiong, “Multi-step data prediction in wireless sensor networks based on one-dimensional CNN and bidirectional LSTM,” *IEEE Access*, vol. 7, pp. 117883–117896, 2019.
- [29] N. Xiong, Y. Yang, H. Jing, and Y. He, “On designing QoS for congestion control service using neural network predictive techniques,” in *Proceedings of the 2006 IEEE International Conference on Granular Computing*, pp. 299–304, Atlanta, GA, USA, May 2006.
- [30] Y. Liu, N. Xiong, Y. Li, K. Xu, J. H. Park, and C. Lin, “A secure model for controlling the hubs in P2P wireless network based on trust value,” *Computer Communications*, vol. 33, no. 8, pp. 997–1004, 2010.
- [31] N. N. Xiong, H. Cheng, S. Hussain, and Y. Qu, “Fault-tolerant and ubiquitous computing in sensor networks,” *International Journal of Distributed Sensor Networks*, vol. 9, no. 10, 2013.
- [32] M. Wu, L. Tan, and N. Xiong, “A structure fidelity approach for big data collection in wireless sensor networks,” *Sensors*, vol. 15, no. 1, pp. 248–273, 2014.
- [33] P. yang, N. Xiong, and R. Jingli, “Data security and privacy protection for cloud storage: a surge,” *IEEE Access*, vol. 8, pp. 131723–131740, 2020.

Research Article

Effective Evolutionary Algorithm for Solving the Real-Resource-Constrained Scheduling Problem

Huu Dang Quoc ¹, Loc Nguyen The ², Cuong Nguyen Doan ³ and Naixue Xiong ⁴

¹Thuong Mai University, Hanoi, Vietnam

²Hanoi National University of Education, Hanoi, Vietnam

³Military Institute of Science and Technology, Hanoi, Vietnam

⁴Northeastern State University, Tahlequah, OK, USA

Correspondence should be addressed to Huu Dang Quoc; huudq@tmu.edu.vn and Loc Nguyen The; locnt@hnue.edu.vn

Received 20 August 2020; Revised 19 September 2020; Accepted 30 September 2020; Published 14 October 2020

Academic Editor: Hongju Cheng

Copyright © 2020 Huu Dang Quoc et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This paper defines and introduces the formulation of the Real-RCPSP (Real-Resource-Constrained Project Scheduling Problem), a new variant of the MS-RCPSP (Multiskill Resource-Constrained Project Scheduling Problem). Real-RCPSP is an optimization problem that has been attracting widespread interest from the research community in recent years. Real-RCPSP has become a critical issue in many fields such as resource allocation to perform tasks in Edge Computing or arranging robots at industrial production lines at factories and IoT systems. Compared to the MS-RCPSP, the Real-RCPSP is supplemented with assumptions about the execution time of the task, so it is more realistic. The previous algorithms for solving the MS-RCPSP have only been verified on simulation data, so their results are not completely convincing. In addition, those algorithms are designed only to solve the MS-RCPSP, so they are not completely suitable for solving the new Real-RCPSP. Inspired by the Cuckoo Search approach, this literature proposes an evolutionary algorithm that uses the function Reallocate for fast convergence to the global extremum. In order to verify the proposed algorithm, the experiments were conducted on two datasets: (i) the iMOPSE simulation dataset that previous studies had used and (ii) the actual TNG dataset collected from the textile company TNG. Experimental results on the iMOPSE simulation dataset show that the proposed algorithm achieves better solution quality than the existing algorithms, while the experimental results on the TNG dataset have proved that the proposed algorithm decreases the execution time of current production lines at the TNG company.

1. Introduction

Scheduling is used to arrange the resources and tasks in many fields, where scheduling algorithms can have an important impact on the effectiveness and cost. In a logistic system, not only are cargo vehicles characterized by the factors of speed and carrying capacity, but also the skill of the driver is the most important factor in determining the quality of transportation. By taking into account all the factors, especially the driver's skill one, scheduling algorithms help manage and coordinate the transport system. An intelligent scheduling algorithm helps managers exploit the maximum potential of resources including vehicle and driver to get the project done.

In the wireless sensor networks, the node scheduling aims at selecting a set of nodes (i.e., sensors) that provide the data service. This scheduling can effectively reduce the number of nodes and messages and at the same time extend the network lifetime [1–3]. The basic goal of Edge Computing [4] is finding the optimal scheduling for extending the Cloud's resources such as servers and routers from remote data centers to the edge of the Cloud where they are closer to users, thus overcoming the bottlenecks issue by cloud computing and providing higher performance.

Solving the MS-RCPSP [5–7] problem is to find out the schedule to execute the project in the shortest possible time without breaking any constraints. In other words, the scheduling algorithm's goal is to find a schedule with the

smallest execution time while meeting any task and resource constraints. In this paper, the term “makespan” will be used to refer to the distance in time that elapses from the start of work to the end of execution time. MS-RCPSP is among the most commonly investigated optimization problems that have received a lot of attention due to their significant role in network resource scheduling and controlling.

MS-RCPSP appears in many practical situations such as logistics and cargo transportation, widely applied to military operations such as sorting missions and determining travel routes [8]. Hosseinian and Baradaran [9] used an evaluation method to make plan decisions with MS-RCPSP. Nazafzad et al. [10] employed a biobjective optimization model for the MS-RCPSP considering shift differential payments and time-of-use electricity tariffs. Their study tried to minimize the cost and the makespan of a given project. Younis and Yang [11] propose the heuristic algorithm to solve a particular case of the MS-RCPSP occurring in grid computing.

However, the MS-RCPSP has one serious defect. What often happens in practice is that a resource with a higher skill level has a shorter processing time. This paper presents a new problem, which is a more practical extension of the MS-RCPSP, called the Real-RCPSP. In other words, Real-RCPSP is a specific case of the MS-RCPSP. In the Real-RCPSP, the processing time depends on the skill level of the resource. The real-life nature of Real-RCPSP comes from production lines at the industry factory, where the higher the skill level of a worker is, the faster he can make the product.

This paper is organized as follows. The next section presents some previous algorithms for solving the Resource-Constrained Scheduling Problem. This section also briefly introduces the Cuckoo Search strategy [12], one of the most widely used metaheuristics. Section 3 formulates the Real-RCPSP. The proposed algorithm (called R-CSM) is described in the fourth section. Section 4 introduces the most important components of R-CSM, consisting of the function Reallocate, the schedule representation, and a novel schedule measurement model. To verify the performance of the proposed algorithm, in Section 5 and Section 6, we arrange the experiments on the iMOPSE dataset and TNG’s dataset, respectively. In these two sections, the experimental results are analyzed to compare the performance of the proposed algorithm R-CSM with the best previous algorithms such as GreedyDO and GA. Finally, Section 7 ends the paper with the conclusion and future works.

2. Related Works

Despite the importance of the Real-RCPSP, no one to the best of our knowledge has studied this problem. This paper is the first work that mentions Real-RCPSP; thus this section introduces the existing algorithms to solve another problem that is close to the Real-RCPSP, namely, MS-RCPSP. In Section 5 and Section 6, these algorithms will be used as reference algorithms in our experiments.

Myszkowski et al. [6] proved that MS-RCPSP is an NP-hard problem, so it is difficult to deal with classical optimization methods. Until now, many different solutions have been introduced for solving the MS-RCPSP; among them,

the most successful metaheuristics are the GA [13] and ACO [14].

In their research, Maghsoudlou et al. [15] and Bibiks et al. [16] applied the Cuckoo Search algorithm to build multirisk project implementation schedules based on three different evaluation objectives. Zhu et al. [17] proposed an evolutionary algorithm based on the multiverse and several other heuristic algorithms.

Myszkowski et al. [6] have built a hybrid algorithm that combined the Difference Evolution and greedy heuristic for managing human and machine resources in factory production projects. The proposed hybrid algorithm tried to minimize makespan and production costs. Besides, iMOPSE [6], a dataset that was generated based on real-world data from the project scheduling problem, was introduced.

As a specific case of the MS-RCPSP problem, Real-RCPSP is researched and applied in many fields of the Internet of Things. Hosseinian and Baradaran [18] proposed a greedy heuristic for maximizing the modularity to find high-quality communities of employees and to arrange them to the tasks based on the founded communities. Younis and Yang [11] introduced a hybrid scheduling algorithm for task arrangement in grid computing environment.

Some other researchers have also studied the new extension problems of the Constrained Project Scheduling Problem and applied them in many fields of science and finance. Polo-Mejia et al. [19] developed a scheduling algorithm to manage nuclear laboratory operations. To solve the problem of the dense sensors in wireless sensor networks, Wan et al. [20] proposed an energy-saving scheduling algorithm, which arranges some redundant sensors into the sleep mode to reduce the data transmission collision and energy dissipation. Guo et al. [21] developed a PSO-based algorithm that acquired better performance than previous approaches in power efficiency. Cheng et al. [22] have formulated another PSO-based algorithm named DPSCA, which is based on the discrete PSO that aims at minimizing the cochannel interference in the network.

Previous studies have also been performed to address other subissues of the Constrained Project Scheduling Problem. Barrios et al. [23] and Javanmard et al. [24] studied the Multiskill Stochastic and Preemptive Scheduling Problem to minimize the execution time and proposed the mathematical models for the project’s resource investment.

Cuckoo Search (CS) algorithm is a metaheuristic introduced by Yang [25] based on the cuckoo bird behavior. Previous algorithms such as Difference Evolution (DE) [26] and Particle Swarm Optimization (PSO) [21] have been proven to be special cases of the Cuckoo Search algorithm. The efficiency of CS has also been shown to be better than those of DE and PSO in some cases [27]. For the above reasons, in this paper, we have built an algorithm inspired by CS.

3. Problem Statement

The Real-RCPSP can be described as follows.

A project represented by a graph $G(V, E)$ has to be realized. Each node of graph G represents a task, while G ’s

arc represents the relationship between 2 tasks (Figure 1). Specifically, the arc (i, j) means that task i has to be finished by the time when task j is started. Each task has an execution time (or duration) that is calculated by subtracting the start time from the end time. The task must be performed continuously from start to finish and must not be stopped at all.

The execution of each task requires some specific resources. Each resource can perform only one task at a time. The task's execution requires several skills, while each resource possesses its own skills; thus not every resource can perform a given task.

The objective of the Real-RCPS is to shorten the project implementation time to the smallest value while not breaking any constraints. A schedule must be found in which execution will minimize execution time while still meeting the task and resource constraints. As mentioned above, the MS-RCPS is proved to be NP-hard [5–7], and no polynomial-time algorithm exists, assuming that $P \neq NP$.

Real-RCPS could be stated by using the following notations:

- (i) C_i : the set of the parents of task i
- (ii) r_i : the set of skills required by task i . A certain resource must have an equal or higher skill level of r_j to perform task j
- (iii) S : the set of all skills; S_i : the set of skills belonging to resource i ; $S_i \subseteq S$
- (iv) t_{jk} : the time it takes the resource that possesses subset of skill S_k to complete task j
- (v) L : the set of resources; L^k : the set of resources that could handle task k ; $L^k \subseteq L$
- (vi) L_i : resource i
- (vii) W : the set of tasks; W^k : the set of tasks that could be performed by resource k , $W^k \subseteq W$
- (viii) W_i : task i
- (ix) B_k, E_k : starting time and ending time of task k
- (x) $A_{u,v}^i$: a Boolean variable; when it equals 1 it means that task u will be executed by resource v at time i ; it equals 0 in other cases
- (xi) h_i : the level of skill i ; g_i : type of skill i
- (xii) r_k : a resource must possess skill r_k to perform task k
- (xiii) m : the period time of the schedule
- (xiv) P : a candidate schedule; P_{all} : the set of candidate schedules
- (xv) $f(P)$: makespan (execution time) of schedule P
- (xvi) y : number of tasks; z : number of resources

Real-RCPS could be defined as follows:

$$\text{minimize } f(P), \quad (1)$$

where

$$f(P) = \max_{i \in T} \{E_i\} - \min_{k \in T} \{B_k\}, \quad (2)$$

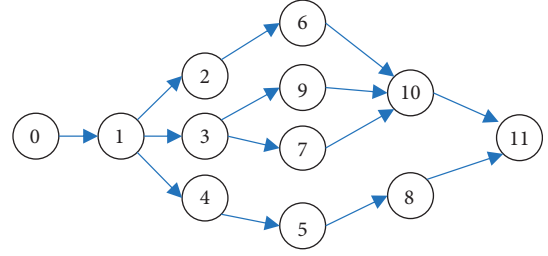


FIGURE 1: The relationship between tasks.

subject to

$$S_k \neq \emptyset, \quad \forall k \in L, \quad (3)$$

$$T_j \geq 0, \quad \forall j \in W, \quad (4)$$

$$E_j \geq 0, \quad \forall j \in W, \quad (5)$$

$$E_i \leq E_j - t_j, \quad \forall j \in W, j \neq 1, i \in C_j, \quad (6)$$

$$\forall i \in W_k \exists r \in S^k: g_r = g_{r_i}, h_r \geq h_{r_i}, \quad (7)$$

$$\forall k \in L, \forall t \in m: \sum_{i=1}^n A_{i,k}^t \leq 1, \quad (8)$$

$$\forall j \in W \exists! t \in m, !k \in L: A_{j,k}^t = 1, \quad (9)$$

$$\text{If } h_i < h_j \text{ then } t_{jk} > t_{ik} \forall (r_i, r_j) \in \{S_k \times S_v\}. \quad (10)$$

Note the following:

- (i) Formulation (6) forced the parent task to must be finished before the start time of the children task
- (ii) Formulation (7) means that, for every task, there is always at least one resource that has enough skill level to handle that task
- (iii) Formulation (8) ensures that each resource (k) can only perform at most one task (j) at any time (t)
- (iv) Constraint (9) aims to restrict each task to only be executed at most one time. Every task must be performed continuously from start to finish and must not be stopped at all.
- (v) Constraint (10) means that the execution time of the higher-skill resources is smaller than the execution time of the lower-skill resources

4. Proposed Algorithm

4.1. Schedule Representation. We represent a schedule as a row that consists of several elements, and the number of elements denotes the number of tasks. Each element of the row represents the resource that will perform the respective task.

4.1.1. *Example 1.* Suppose that we have 10 tasks $W = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ and 3 resources $L = \{1, 2, 3\}$. Assume the following:

- (i) $S_k = S, \forall k \in L$; every resource has an equal skill set
- (ii) $L^k = L, \forall k \in W$; any resource can perform any task
- (iii) The execution times of the tasks are presented in Table 1

We also assume that the constraint to prioritize the performance of the tasks is shown in Figure 1, specifically:

- (i) Task 1 has to be performed firstly
- (ii) Task 6 has to be performed after task 2
- (iii) Tasks 7 and 9 have to be performed after task 3
- (iv) Task 5 has to be performed after task 4
- (v) Task 8 has to be performed after task 5
- (vi) Task 10 has to be performed after tasks 6, 7, and 9

With the above assumptions and constraints, a possible schedule is shown in Figure 2. Table 2 shows how that schedule assigns 3 resources to perform 10 tasks in detail.

As described in Table 2, resource 1 executes task 2 and task 6; resource 2 handles tasks 1, 3, 5, and 8; resource 3 executes tasks 4, 7, 9, and 10.

4.2. *Measurement Model.* Cuckoo Search algorithm is an optimization scheme dealing with real functions such as the Gaussian probability distribution function, whereas Real-RCPSp is the optimization problem of discrete functions. Therefore, in order to apply the Cuckoo Search algorithm to the Real-RCPSp, it is necessary to build a model for schedules measuring. The following will present our proposed measurement model in detail:

- (i) $V = (v_1, v_2, \dots, v_n)$ is called “unit vector,” where $v_i = 100/(z_i - 1)$; z_i is the number of resources that possess set of skills q_i .
- (ii) Vector $K = \{k_1, k_2, \dots, k_n\}$ is the distance between schedule $P = \{p_1, p_2, \dots, p_n\}$ and schedule $Q = \{q_1, q_2, \dots, q_n\}$. This leads to $K = P - Q$.

Meanwhile, if schedule $Q = \{q_1, q_2, \dots, q_n\}$ is added with a difference $K = \{k_1, k_2, \dots, k_n\}$, schedule $P = \{p_1, p_2, \dots, p_n\}$ is obtained, where we have the following:

- (iii) $p_i = \text{position}(\text{round}(q_i + k_i))$ and position (i) presents the respective resource
- (iv) $k_i = v_i \times (\text{order}(p_i) - \text{order}(q_i))$

$\text{order}(p_i)$: the place of p_i in the L_i

4.2.1. *Example 2.* Suppose that $L^1 = \{L_1, L_3, L_4, L_9, L_{10}\}$. We have $z_1 = 5$; $v_1 = 100/(5 - 1) = 25$.

TABLE 1: Execution times of the tasks.

Task	W_1	W_2	W_3	W_4	W_5	W_6	W_7	W_8	W_9	W_{10}
Execution time	3	2	3	4	3	4	3	6	2	6

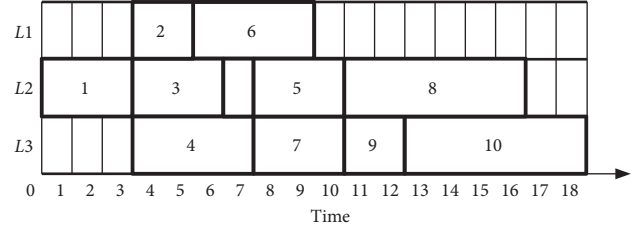


FIGURE 2: A possible schedule.

TABLE 2: The assignment of a possible schedule.

Task	W_1	W_2	W_3	W_4	W_5	W_6	W_7	W_8	W_9	W_{10}
Resource	L_2	L_1	L_2	L_3	L_2	L_1	L_3	L_2	L_3	L_3

TABLE 3: The order of resources.

Order	0	1	2	3	4
Resource	L_1	L_3	L_4	L_9	L_{10}
Resource	L_6	L_7	L_8	L_{10}	

Similarly, suppose that $L^1 = \{L_6, L_7, L_8, L_{10}\}$. We have $z_2 = 4$; $v_2 = 100/(4 - 1) = 33.33$.

Table 3 depicts the order of resources.

Schedule $P = (1, 8)$ and schedule $Q = (4, 7)$ are shown in Table 4.

Consider the distance $K = P - Q = (k_1, k_2)$ where $k_1 = v_1 \times \text{abs}(\text{order}(p_1) - \text{order}(q_1)) = 25 \times \text{abs}(0 - 2) = 25$.

$k_2 = v_2 \times \text{abs}(\text{order}(p_2) - \text{order}(q_2)) = 33.33 \times \text{abs}(2 - 1) = 33.33$.

This leads to $K = (50, 33.33)$.

Given $K_P = (0, 66.66)$, we have $Z = P + K \rightarrow K_P + K \rightarrow (4, 10)$ (Table 5).

4.3. *Proposed Algorithm R-CSM.* The proposed Algorithm 1 R-CSM is represented as follows.

f is objective function.

Note that, in line number 16, function *Reallocate()* improves the quality of b_plan , as analyzed in the next subsection.

4.4. *Function Reallocate.* c_plan is the most appropriate feasible schedule until now. L_b is the last resource to finish. Function *Late()* will find out the value of R_b . *Size()* is the size of a set or an array. $N_makespan$ is execution time of the new resource-task arrangement

TABLE 4: The assignment of schedules P and Q .

Schedule	Task	
	1	2
P	L_1	L_8
Q	L_4	L_7

TABLE 5: Measurement value.

Task	1	2
P	L_1	L_8
K_p	0	66.66
K	50	33.33
$K_p + K$	50	99.99
$Z = K_p + K$	L_4	L_{10}

```

input: maxGeneration
      iMOPSE datasets
output: makespan of project
(1) Begin
(2)   $t \leftarrow 0$ 
(3)  Size  $\leftarrow$  number of individuals (i.e. possible schedules)
(4)   $p(t) \leftarrow$  the first population
(5)   $f(t) \leftarrow$  the fitness, b_plan(bestnest), makespan
(6)  pa = 0.25
(7)  While ( $t < max\_gen$ )
(8)    n_plan  $\leftarrow$  create new nest by Lévy Flight
(9)    r_plan  $\leftarrow$  Select random nest from  $P(t)$ 
(10)   If ( $f(n\_plan) < (r\_plan)$ )
(11)     r_plan = n_plan
(12)   End if
(13)    $P(t) \leftarrow$  Remove pa worst nest and replace by new nests, new nests created by Lévy Flight
(14)    $F(t) \leftarrow$  the fitness, b_plan, makespan
(15)   b_plan  $\leftarrow$  Reallocate(b_plan)//schedule b_plan is improved by the//function Reallocate(), which is described in the next
      subsection in details.
(16)    $t \leftarrow t + 1$ 
(17) End while
(18) return makespan
(19) End

```

ALGORITHM 1: R-CSM algorithm.

Line 12 and line 13 show that the new schedule (n_plan) is always equal to or better than the old schedule (c_plan) in terms of the makespan.

The function Reallocate() (Figure 3) generates the new schedule from the best schedule, so it inherits and promotes the advantages of the current population (Algorithm 2).

4.4.1. *Example 3.* Suppose that $W = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$; $L = \{1, 2, 3\}$.

We also assume that resource 1 can handle tasks 1, 2, 3, 4, 6, 8, 9, and 10; resource 2 can execute tasks 1, 3, 7, and 9; resource 3 can perform tasks 1, 4, 5, 8, 9, and 10.

The constraint regarding the task order is illustrated in Figure 1, and the task's execution time is shown in Table 1.

Table 6 depicts the schedule P ; its makespan is equal to 18 as shown in Figure 4 in detail.

The function Reallocate uses schedule P as the input and arranges task 9 to resource 1 (see Table 7) instead of resource 3 as before.

The results point out that function *Reallocate* decreases the makespan from 18 to 17, as described in Figure 4.

5. Simulation with iMOPSE Dataset

5.1. *Simulation Settings.* To verify the performance of the R-CSM, the simulations are conducted by using iMOPSE dataset [6], which has been used by previous studies to examine algorithms such as GreedyDO and GA [28]. iMOPSE's instances have the following fields:

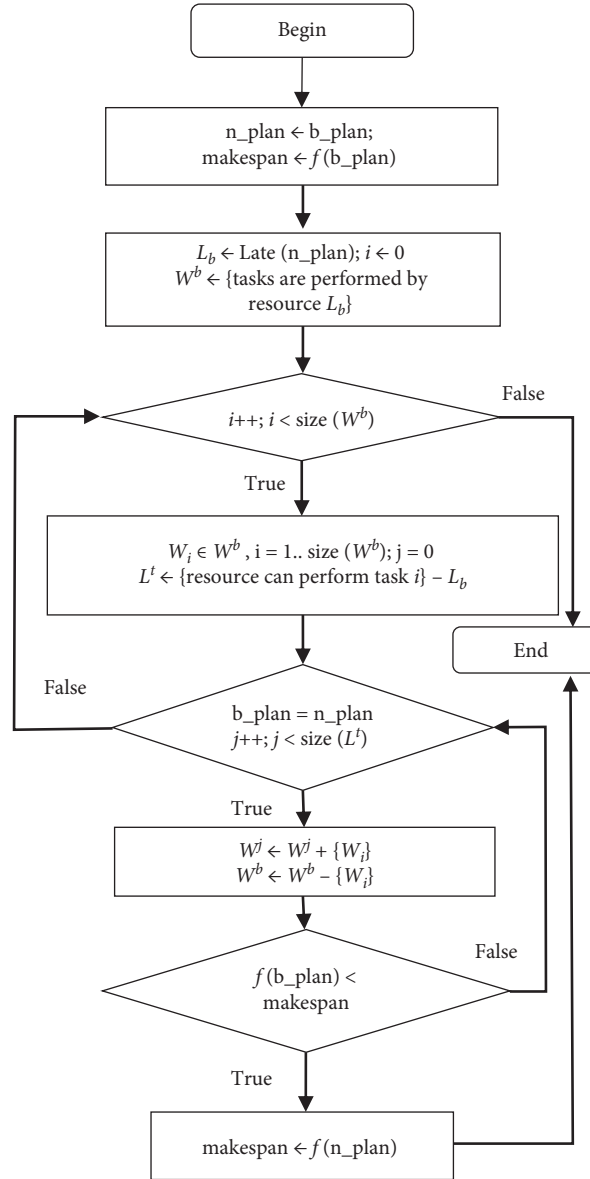


FIGURE 3: Function Reallocate.

- (i) Number of tasks and resources
- (ii) The constraint regarding the task order
- (iii) Set of resource's skills

- (iii) The program execution process consists of 50,000 generations ($N_g = 50,000$)
- (iv) Each instance was repeatedly executed 30 times

This paper arranges the simulations on iMOPSE's instances listed in Table 8. All of our simulations were run on a machine with Intel® Core i7 CPU at 2.2 GHz, 6 GB RAM, running Windows 10. Our R-CSM algorithm was programmed in Matlab. Simulation results are described in Table 9.

Each simulation was set up with parameters as follows:

- (i) Dataset: 30 iMOPSE's instances that are described above
- (ii) Number of individuals in population $N_p = 100$

5.2. Simulation Results. To show the efficiency of the proposed algorithm, we compare R-CSM with two existing algorithms, which are GreedyDO and GA [28]. Myszkowski did not provide the tool for GreedyDO; thus Table 9 just lists the best value of the algorithm GreedyDO that was published in the author's literature. Meanwhile algorithm GA is reprogrammed using the GARunner, the tool provided by the authors in [6, 28]; thus Table 9 lists the average value, the best value, and the standard deviation value of the algorithm GA's makespan.

```

Input: b_plan//the best feasible schedule that the algorithm has found until now
Output://the feasible schedule which is better than the input
(1) Begin
(2) makespan = f(b_plan)
(3) n_plan = b_plan;//the best resource-task assignment plan so far
(4)  $L_b \leftarrow \text{Late}(n\_plan)$ //the resource finish its execution latest
(5)  $W_b \leftarrow$  set of tasks is performed by resource  $L_b$ 
(6) For  $i = 1$  to size( $W_b$ )//examine the set of tasks performed by resource  $L_b$ 
(7)    $W_i = W_b[i]$ ;
(8)    $L^i \leftarrow L - L_b$   $L^i \leftarrow L - L_b$ //set of resource can perform the task  $i$  except  $L_b$ 
(9)   For  $j = 1$  to size( $L^i$ )
(10)     $W^j = W^j + \{W_i\}$ //task  $i$  will perform the resource  $L_j$ 
(11)     $W^b = W^b + \{W_i\}$ //task  $i$  is eliminated from  $L_b$ 
(12)    n_makespan = f(n_plan)
(13)    If n_makespan < makespan
(14)     Makespan = n_makespan
(15)     Return b_best;
(16)    End if
(17)   b_plan = n_plan;
(18)   End for
(19) End for
(20) Return n_plan
(21) End Function
    
```

ALGORITHM 2: Function Reallocate.

TABLE 6: Resource-task assignment of P .

Task	W_1	W_2	W_3	W_4	W_5	W_6	W_7	W_8	W_9	W_{10}
Resource	L_1	L_1	L_2	L_3	L_3	L_1	L_2	L_1	L_3	L_3

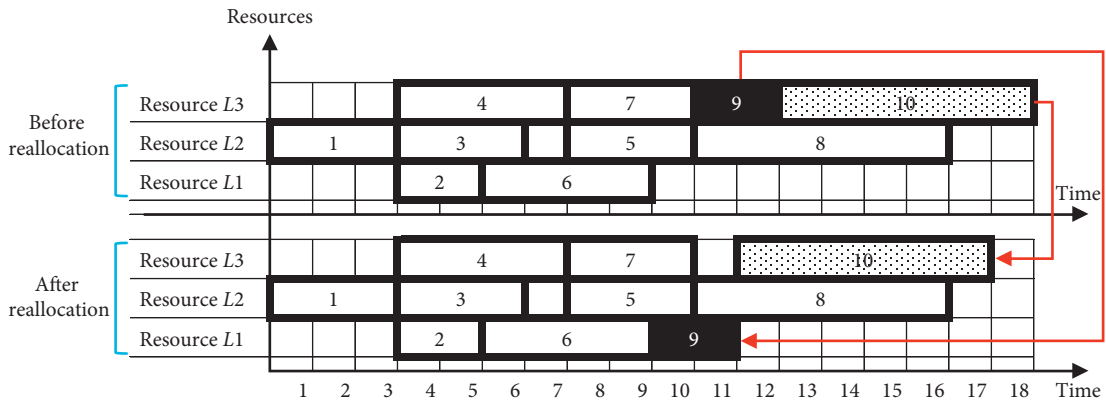


FIGURE 4: The schedule changes as a result of the function Reallocate.

TABLE 7: Resource-task assignment of new P .

Task	W_1	W_2	W_3	W_4	W_5	W_6	W_7	W_8	W_9	W_{10}
Resource	L_1	L_1	L_2	L_3	L_3	L_1	L_2	L_1	L_1	L_3

Table 9 and Figure 5 demonstrated that the makespan of the R-CSM's schedules is smaller than the makespan of GreedyDO and GA. The comparison between algorithms is discussed as follows in detail:

- (i) Compared with the original CS, the R-CSM algorithm is equipped with function Reallocate, which makes R-CSM capable of fast convergence. This ability is clearly demonstrated by the comparison of

TABLE 8: d36 iMOPSE dataset.

Dataset instance	Tasks	Resources	Precedence relations	Skills
100_5_22_15	100	5	22	15
100_5_46_15	100	5	46	15
100_5_48_9	100	5	48	9
100_5_64_15	100	5	64	15
100_5_64_9	100	5	64	9
100_10_26_15	100	10	26	15
100_10_47_9	100	10	47	9
100_10_48_15	100	10	48	15
100_10_64_9	100	10	64	9
100_10_65_15	100	10	65	15
100_20_22_15	100	20	22	15
100_20_46_15	100	20	46	15
100_20_47_9	100	20	47	9
100_20_65_15	100	20	65	15
100_20_65_9	100	20	65	9
200_10_128_15	200	10	128	15
200_10_50_15	200	10	50	15
200_10_50_9	200	10	50	9
200_10_84_9	200	10	84	9
200_10_85_15	200	10	85	15
200_20_145_15	200	20	145	15
200_20_54_15	200	20	54	15
200_20_55_9	200	20	55	9
200_20_97_15	200	20	97	15
200_20_97_9	200	20	97	9
200_40_133_15	200	40	133	15
200_40_45_15	200	40	45	15
200_40_45_9	200	40	45	9
200_40_90_9	200	40	90	9
200_40_91_15	200	40	91	15

R-CSM with the previous most powerful algorithms. R-CSM's best value is smaller than GreedyDO from 21% to 85% and faster than GA from 6% to 33%. The average value of the R-CSM's makespan is better than the GA from 6% to 33%.

- (ii) Thanks to the function Reallocate and the proposed measurement model, the process of finding the optimal schedule of the R-CSM is not only fast but also stable. This is demonstrated by the experimental results in Table 9. The total value of R-CSM's standard deviation is 92.52 only, whereas the total value of GA's standard deviation is equal to 180. This result shows that the R-CSM algorithm is more stable than the GA algorithm.

6. Experiment with TNG Dataset

6.1. Experimental Setting. In general, the major disadvantage of verifying on simulation datasets, such as iMOPSE, is that sometimes the results do not match what is actually happening. In order to make the experiment more convincing, we have collected and used the dataset of Investment and Trading Joint Stock Company (TNG) [29]. At TNG textile factory, the dataset construction is carried out as follows:

TABLE 9: Makespan of algorithms (in hour).

Dataset instance	GreedyDO	GA			R-CSM		
		Avg	Best	Std	Avg	Best	Std
100_5_22_15	630	524	517	5	485	484	1.41
100_5_46_15	693	587	584	5	541	538	2.55
100_5_48_9	779	535	528	10	493	490	2.16
100_5_64_15	640	530	527	2	496	490	4.55
100_5_64_9	597	521	508	10	479	474	3.68
100_10_26_15	370	294	292	2	238	237	0.94
100_10_47_9	549	299	296	3	256	253	2.36
100_10_48_15	344	282	279	3	245	242	2.25
100_10_64_9	533	305	296	7	249	243	5.32
100_10_65_15	426	290	286	5	247	245	1.25
100_20_22_15	353	169	163	6	127	124	3.4
100_20_46_15	394	207	197	7	167	164	2.85
100_20_47_9	390	186	185	0	146	143	1.89
100_20_65_15	310	243	240	2	213	210	2.05
100_20_65_9	408	187	181	5	133	128	4.5
200_10_128_15	780	583	577	5	471	468	2.94
200_10_50_15	763	577	553	17	489	485	3.3
200_10_50_9	817	589	585	5	484	484	0.47
200_10_84_9	999	583	567	11	509	505	3.91
200_10_85_15	706	555	549	5	476	474	1.32
200_20_145_15	480	328	326	2	244	240	3.48
200_20_54_15	488	385	363	21	261	257	3.19
200_20_55_9	999	318	312	4	251	248	2.87
200_20_97_15	680	438	424	10	335	334	1.89
200_20_97_9	816	326	321	6	249	244	3.42
200_40_133_15	512	222	215	6	154	148	4.19
200_40_45_15	616	210	201	6	165	161	3.68
200_40_45_9	821	213	209	3	159	152	8.06
200_40_90_9	963	215	211	3	152	148	3.72
200_40_91_15	519	205	200	3	141	135	4.92

- (i) The company TNG contracts with business partners, whereby each order corresponds to a product sample with a large quantity
- (ii) A given order will be performed by a subset of employees
- (iii) A product consists of several components, and each component takes an execution time
- (iv) The skills of each worker are evaluated based on that worker's rank

Conducting experiments on a simulation dataset is always convenient because the parameters of the dataset are set by the experimenter; therefore these parameters are completely consistent with the problem formulation.

In contrast, the parameters of a real dataset are factory-defined, so they are not compatible with the conventions in the problem formulation. For this reason, before conducting experiments with TNG's dataset, the parameters of this dataset need to be converted to a format that matches the Real-RCSP formulation.

This conversion is conducted as follows:

- (i) Order is demonstrated by the project
- (ii) Product's stage is depicted by a task
- (iii) An employee is depicted by a resource

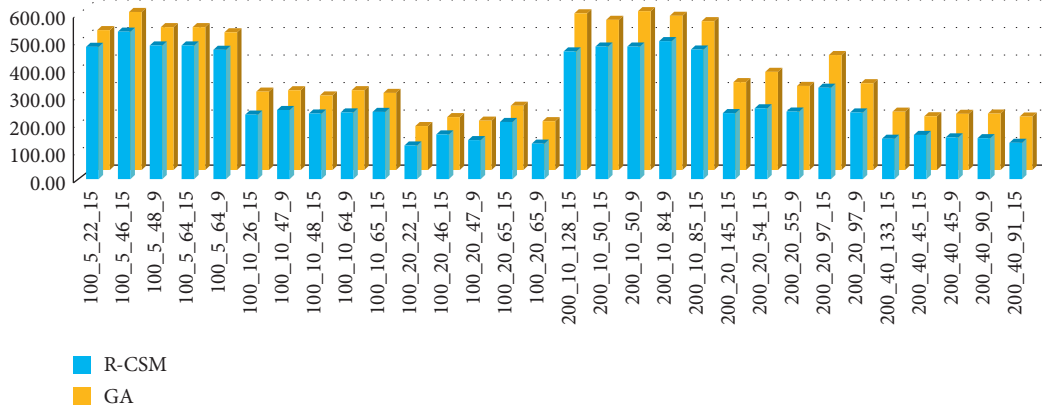


FIGURE 5: The comparison of performance between R-CSM and GA.

TABLE 10: TNG dataset.

Dataset instance	Name	Tasks	Resources	Precedence relations	Skill levels	Project time
71_37_1026_1	TNG1	71	37	1026	6	409
71_39_1026_1	TNG2	71	39	1026	6	325
71_41_1026_1	TNG3	71	41	1026	6	296
71_45_1026_1	TNG4	71	45	1026	6	392
137_37_1894_1	TNG5	137	37	1894	6	1174
137_39_1894_1	TNG6	137	39	1894	6	1052
137_41_1894_1	TNG7	137	41	1894	6	871
137_45_1894_1	TNG8	137	45	1894	6	996

TABLE 11: Makespan of GreedyDO, GA, and R-CSM (in hour).

Dataset instance	TNG	GreedyDO	GA	R-CSM
TNG1	409	236	201	166
TNG2	325	243	198	165
TNG3	296	258	212	168
TNG4	392	248	176	175
TNG5	1174	972	751	710
TNG6	1052	963	791	715
TNG7	871	834	810	727
TNG8	996	906	720	677

- (iv) Employee's grade is depicted by the resource's skill level
- (v) The manufacture sequence is denoted by the task's relationship
- (vi) The execution time of the order is demonstrated by the makespan

The TNG dataset is described in Table 10. Experiment setting is as follows:

- (i) Dataset: 8 TNG's instances that are listed in Table 10
- (ii) Number of individuals in population $N_p = 100$
- (iii) The program execution process consists of 50,000 generations ($N_g = 50,000$)
- (iv) Each instance was repeatedly executed 35 times

6.2. Experimental Results. The experiments in this section were conducted on the TNG dataset to prove that the proposed algorithm is more efficient than existing algorithms not only when they are operating on the simulated dataset such as iMOPSE but also when operating on the actual dataset.

Experimental results (listed in Table 11) demonstrated that the proposed algorithm R-CSM is not only more efficient than previous algorithms such as GreedyDO and GA but also more efficient than the actual production plan at the TNG factory, which is presented in column TNG.

As depicted in Table 11, compared to the execution time of the actual production plan at the factory TNG, the GA and GreedyDO algorithms reduce the makespan by 7%–55% and 4%–42%, respectively, while R-CSM has the best results of 17%–59%.

TABLE 12: Experimental results with Real-RCPSP.

Dataset instance	TNG (A)	MS-RCPSP (1)		Real-RCPSP (2)		(2) vs. (1)	
		Best	vs. (A) (%)	Best	vs. (A) (%)	Hours	%
TNG1	409	166	59	131	68	35	21.1
TNG2	325	165	49	133	59	32	19.4
TNG3	296	168	43	132	55	36	21.4
TNG4	392	175	55	127	68	48	27.4
TNG5	1174	710	40	572	51	138	19.4
TNG6	1052	715	32	626	40	89	12.4
TNG7	871	727	17	569	35	158	21.7
TNG8	996	677	32	560	44	117	17.3

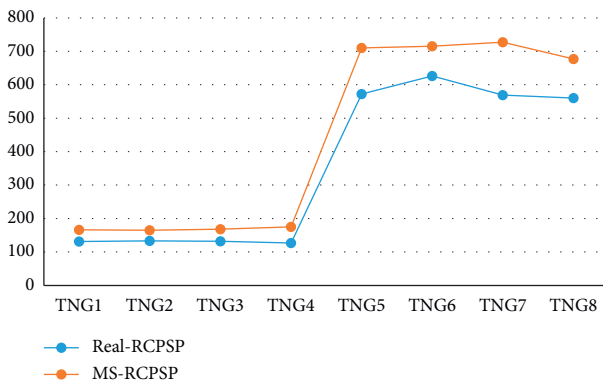


FIGURE 6: R-CSM to solve MS-RCPSP and Real-RCPSP.

As the best current evolutionary algorithms, GreedyDO and GA are both better than the actual production plan at the factory TNG. However, neither of these algorithms is as good as the proposed algorithm R-CSM. In experiments on the TNG datasets, R-CSM lessens the makespan from 17% to 59% compared to the current factory schedule. To sum up, the proposed algorithm R-CSM has been proven to be overall more effective compared to existing approaches such as GreedyDO, GA, and the current factory schedule.

Applying the R-CSM algorithm on the data from table Table 11 to solve the Real-RCPSP, we get the results shown in Table 12.

Experimental results in Table 12 show that, thanks to the application of the R-CSM algorithm to the actual problem with data on the textile production of TNG, the production time is reduced from 12.4% to 27%. These results also show that the greater the difference between the skill levels of workers, the more efficient the R-CSM algorithm.

Figure 6 shows the effectiveness of algorithm R-CSM when it is applied to solve the MS-RCPSP and the Real-RCPSP on the textile dataset of TNG textile company.

7. Conclusion

This article aims to announce and survey Real-RCPSP, a new combinatorial optimization problem that appears in many fields such as Edge Computing, industrial production, and IoT systems. The new problem is stated, and then a new algorithm named R-CSM is proposed. Inspired by the Cuckoo Search

strategy, the proposed algorithm has been upgraded by using function Reallocate; thus it achieved high performance.

The experimental results show that the proposed algorithm R-CSM is better than the previous algorithms such as GreedyDO and GA at 21%–85% and 6%–33%, respectively. At the same time, the proposed algorithm converged to the optimal solution faster than previous approaches.

In the near future, we are going to continue researching on Real-RCPSP in order to improve the solution quality and the speed of the convergence. Multifactorial optimization seems to be one of the promising approaches to this problem.

Data Availability

The paper uses the standard iMOPSE dataset to test the efficiency of the algorithm. This dataset is publicly available at <http://imopse.ii.pwr.wroc.pl/> and is free of charge. In addition, the authors also tested the algorithm with TNG's garment manufacturing dataset. They have obtained permission from TNG to use these data.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors greatly acknowledge TNG Corporation (434/1 Bac Kan, Thai Nguyen, Vietnam) [29] for their cooperation and for allowing them to use their dataset regarding the production line.

References

- [1] H. Cheng, Z. Su, N. Xiong, and Y. Xiao, "Energy-efficient node scheduling algorithms for wireless sensor networks using markov random field model," *Information Sciences*, vol. 329, no. 2, pp. 461–477, 2016.
- [2] C. Lin, N. Xiong, J. H. Park, and T.-h. Kim, "Dynamic power management in new architecture of wireless sensor networks," *International Journal of Communication Systems*, vol. 22, no. 6, pp. 671–693, 2009.
- [3] C. Lin, Y. X. He, and N. Xiong, "An energy-efficient dynamic power management in wireless sensor networks," in *Proceedings of Fifth International Symposium on Parallel and Distributed Computing*, IEEE, Timisoara, Romania, July 2006.

- [4] Y. Zhou, D. Zhang, and N. Xiong, "Post-cloud computing paradigms: a survey and comparison," *Tsinghua Science and Technology*, vol. 22, no. 6, pp. 714–732, 2017.
- [5] R. Klein, *Scheduling of Resource Constrained Project*, Springer Science Business Media, New York, NY, USA, 2000.
- [6] Myszkowski, B. Paweł, E. Marek, Skowroński, and S. Krzysztof, "A new benchmark dataset for multi-skill resource-constrained project scheduling problem," in *Proceedings of 2015 Federated Conference on Computer Science and Information Systems (FedCSIS)*, IEEE, Lodz, Poland, September 2015.
- [7] J. Błazewicz, J. K. Lenstra, and A. H. G. Rinnooy Kan, "Scheduling subject to resource constraints: classification and complexity," *Discrete Applied Mathematics*, vol. 5, pp. 11–24, 1983.
- [8] H. Li and K. Womer, "A decomposition approach for shipboard manpower scheduling," *Military Operations Research*, vol. 14, no. 3, pp. 1–24, 2009.
- [9] A. H. Hosseinian and V. Baradaran, "An evolutionary algorithm based on a hybrid multi-attribute decision making method for the multi-mode multi-skilled resource-constrained project scheduling problem," *Journal of Optimization in Industrial Engineering*, vol. 12, no. 2, pp. 155–178, 2019.
- [10] H. Najafzad, H. Davari-Ardakani, and R. Nemati-Lafmejani, "Multi-skill project scheduling problem under time-of-use electricity tariffs and shift differential payments," *Energy*, vol. 168, pp. 619–636, 2019.
- [11] M. T. Younis and S. Yang, "Hybrid meta-heuristic algorithms for independent job scheduling in grid computing," *Applied Soft Computing*, vol. 72, pp. 498–517, 2018.
- [12] X. S. Yang and S. Deb, "Cuckoo search via Lévy flights," in *Proceedings of World Congress on Nature & Biologically Inspired Computing (NaBIC 2009)*, IEEE Publications, Coimbatore, India, pp. 210–214, December 2009.
- [13] W. Wu, A. R. Simpson, and H. R. Maier, "Accounting for greenhouse gas emissions in multiobjective genetic algorithm optimization of water distribution systems," *Journal of Water Resources Planning and Management*, vol. 136, no. 2, pp. 146–155, 2010.
- [14] W. Deng, J. Xu, and H. Zhao, "An improved ant colony optimization algorithm based on hybrid strategies for scheduling problem," *IEEE Access*, vol. 7, pp. 20281–20292, 2019.
- [15] H. Maghsoudlou, B. Afshar-Nadjafi, and S. T. Akhavan Niaki, "Multi-skilled project scheduling with level-dependent rework risk; three multi-objective mechanisms based on cuckoo search," *Applied Soft Computing*, vol. 54, pp. 46–61, 2017.
- [16] K. Bibiks, Y.-F. Hu, J.-P. Li, P. Pillai, and A. Smith, "Improved discrete cuckoo search for the resource-constrained project scheduling problem," *Applied Soft Computing*, vol. 69, pp. 493–503, 2018.
- [17] L. Zhu, J. Lin, and Z.-J. Wang, "A discrete oppositional multi-verse optimization algorithm for multi-skill resource constrained project scheduling problem," *Applied Soft Computing*, vol. 85, Article ID 105805, 2019.
- [18] A. H. Hosseinian and V. Baradaran, "Detecting communities of workforces for the multi-skill resource-constrained project scheduling problem: a dandelion solution approach," *Journal of Industrial and Systems Engineering*, vol. 12, pp. 72–99, 2019.
- [19] O. Polo-Mejía, C. Artigues, P. Lopez, and V. Basini, "Mixed-integer/linear and constraint programming approaches for activity scheduling in a nuclear research facility," *International Journal of Production Research*, pp. 1–18, 2019.
- [20] R. Wan, N. Xiong, and N. T. Loc, "An energy-efficient sleep scheduling mechanism with similarity measure for wireless sensor networks," *Human-centric Computing and Information Sciences*, vol. 8, p. 18, 2018.
- [21] W. Guo, N. Xiong, A. V. Vasilakos, G. Chen, and C. Yu, "Distributed k-connected fault-tolerant topology control algorithms with PSO in future autonomic sensor systems," *International Journal of Sensor Networks*, vol. 12, no. 1, pp. 53–62, 2012.
- [22] H. Cheng, N. Xiong, A. V. Vasilakos, L. Tianruo Yang, G. Chen, and X. Zhuang, "Nodes organization for channel assignment with topology preservation in multi-radio wireless mesh networks," *Ad Hoc Networks*, vol. 10, no. 5, pp. 760–773, 2012.
- [23] A. Barrios, F. Ballestín, and V. Valls, "A double genetic algorithm for the MRCPSP/max," *Computers & Operations Research*, vol. 38, no. 1, pp. 33–43, 2011.
- [24] S. Javanmard, B. Afshar-Nadjafi, and S. T. Akhavan Niaki, "Preemptive multi-skilled resource investment project scheduling problem: mathematical modelling and solution approaches," *Computers & Chemical Engineering*, vol. 96, pp. 55–68, 2017.
- [25] X. S. Yang, *Nature-Inspired Optimization Algorithms*, Elsevier, London, UK, 1st edition, 2014.
- [26] M. I. Solihin and M. F. Zaili, "Performance comparison of Cuckoo search and differential evolution algorithm for constrained optimization," *International Engineering Research and Innovation Symposium (IRIS)*, vol. 160, no. 1, 7 pages, 2016.
- [27] M. A. Adnan and M. A. Razzaque, "A comparative study of particle swarm optimization and cuckoo search techniques through problem-specific distance function," in *Proceedings of International Conference on Information and Communication Technology (ICoICT)*, Bandung, Indonesia, March 2013.
- [28] P. B. Myszkowski, Ł. P. Olech, M. Laszczyk, and M. E. Skowroński, "Hybrid differential evolution and greedy algorithm (DEGR) for solving multi-skill resource-constrained project scheduling problem," *Applied Soft Computing*, vol. 62, pp. 1–14, 2018.
- [29] "TNG Investment and trading joint stock company," <http://www.tng.vn>.

Research Article

Urban Traffic Flow Forecast Based on FastGCRNN

Ya Zhang,¹ Mingming Lu ,¹ and Haifeng Li²

¹School of Computer Science and Engineering, Central South University, Changsha 410083, Hunan, China

²School of Geosciences and Info-Physics, Central South University, Changsha 410083, Hunan, China

Correspondence should be addressed to Mingming Lu; mingminglu@csu.edu.cn

Received 23 June 2020; Revised 9 August 2020; Accepted 15 September 2020; Published 27 September 2020

Academic Editor: Naixue Xiong

Copyright © 2020 Ya Zhang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Traffic forecasting is an important prerequisite for the application of intelligent transportation systems in urban traffic networks. The existing works adopted RNN and CNN/GCN, among which GCRN is the state-of-the-art work, to characterize the temporal and spatial correlation of traffic flows. However, it is hard to apply GCRN to the large-scale road networks due to high computational complexity. To address this problem, we propose abstracting the road network into a geometric graph and building a Fast Graph Convolution Recurrent Neural Network (FastGCRNN) to model the spatial-temporal dependencies of traffic flow. Specifically, we use FastGCN unit to efficiently capture the topological relationship between the roads and the surrounding roads in the graph with reducing the computational complexity through importance sampling, combine GRU unit to capture the temporal dependency of traffic flow, and embed the spatiotemporal features into Seq2Seq based on the Encoder-Decoder framework. Experiments on large-scale traffic data sets illustrate that the proposed method can greatly reduce computational complexity and memory consumption while maintaining relatively high accuracy.

1. Introduction

Traffic forecasting using timely information provided by Internet of Things technology (IoT) is an important prerequisite for the application of intelligent transportation system (ITS) [1] in urban traffic networks, because an accurate and efficient prediction model can be used for travellers to select high-quality reference routes, maximize the utilization of road networks, and provide a basis for the reasonable planning of urban construction departments. However, along with worldwide urbanization, urban road networks have been expanded significantly [2], which brings challenges for traffic forecasting because the corresponding computation complexity will greatly increase due to the expanded road networks [3].

This paper mainly studies the problem of urban traffic forecasting based on the Internet of Things technology (IoT) in large urban road traffic networks. This problem is how to use historical traffic flow data to predict traffic flow data in future timestamps in large urban road traffic networks. In the literature, there has been plenty of studies in traffic forecasting, including traffic volume, taxi pick-ups, and

traffic in/out flow volume. Initially, numerous statistical based methods, such as Historical Average (HA) [4], Time Series [5], K Nearest Neighbors Algorithm (KNN) [6], and Kalman Filter [7], have been proposed to predict road traffic. However, these models are generally suitable for relatively stable traffic flow, which cannot well reflect the temporal correlation of traffic flow data, nor can they reflect the real-time nature of traffic flow. In order to solve the unstable characteristics of traffic flow data, ARIMA [8] and its variants [9, 10] are used in this field [11]. Although these studies show that the prediction can be improved by considering various other factors, they are still unable to capture the complex nonlinear spatiotemporal correlation. The latest advances in deep learning enable researchers to model complex nonlinear relationships and show promising results in multiple fields. This success has inspired many attempts to use deep learning technology in traffic flow prediction. Recent studies have proposed the use of improved LSTM [12] and GRU [13] to predict traffic flow. Furthermore, considering the influence of spatial structure on the traffic flow of different roads, Li et al. [14, 15] proposed modeling the traffic volume of the city as an image and partitioning the

city map (the image) into a large number grid. Within each grid cell, the traffic volume within a period of time can be regarded as a pixel value. Based on that, Li adopted ConvLSTM [16] to model the spatial-temporal correlation among traffic flows, where the convolution operation and the LSTM unit are utilized to model spatial and temporal correlation, respectively. However, the conversion of traffic flow into images loses the spatial topology of urban roads. Li et al. [17] modeled the traffic flow as a diffusion process on a directed graph and captured the spatial dependency using bidirectional random walks on the graph and the temporal dependency using the Encoder-Decoder architecture with scheduled sampling. Seo et al. [18] used GCN [19–21] to extract the spatial topology of the traffic network and RNN to find dynamic patterns to optimize traffic forecasting. However, GCN suffers from the scalability issue, because it requires a lot of space to maintain the entire graph and embed each node in memory [22–26], and it has a very high computational complexity [27].

In order to solve the above problem, we propose forming the road network into a geometric graph and constructing a spatiotemporal graph convolution network based on the abstract graph to capture the spatiotemporal features of traffic flow for prediction. We propose using GCN as the spatial topology extractor of the model and applying the sampling method [28–30] to GCN. The method can put the nodes in the graph into the model in batches, sample the neighbors of the nodes in each batch, extract the nodes that have a greater impact on the nodes in this batch, and perform convolution operations, which greatly reduces the calculation complexity and memory. FastGCN can effectively process the large graph by importance sampling so that memory overflow is not easy to occur. Then, we further combine GRU to extract temporal features to achieve the extraction of spatiotemporal features of traffic flow. Finally, we embed spatiotemporal features into Seq2Seq [31] based on Encoder-Decoder framework for prediction.

2. Problem Analysis

Urban traffic flow prediction is based on historical traffic flow sequences, which are highly time-varying, nonlinear, and uncertain. The traffic flow in the road network usually has the following temporal characteristics [32]:

- (a) *Periodicity*. Traffic flows change periodically. The time series of traffic flow usually presents a wavy or oscillatory fluctuation around the long-term trend.
- (b) *Trend and Trend Variability* [33]. The time series of traffic flow shows a regular change trend. It will not change randomly, but it will continuously change with time. For example, from spring to summer, the traffic volume of the morning peak will gradually advance. Present a trending change.
- (c) *Continuity*. Traffic flow has continuity in time; that is, there is a correlation between the value of traffic flow at different times, especially in adjacent time periods.

At a certain time, traffic flow also has some spatial characteristics, such as the impact of traffic flow upstream and downstream of the road on the current road, and the rules of speed limit and traffic flow limit of the same level of road.

In view of these two main influence factors, especially considering the large scale of the road network [34–39], which requires a lot of time for spatial calculation, this paper proposes the Fast Graph Convolution Recurrent Neural Network (FastGCRNN).

It uses recurrent neural network to capture the long-term temporal dependency of traffic flow and the graph convolution neural network (GCN) to capture the spatial correlation among roads in different geographical locations. At the same time, importance sampling is applied to GCN to reduce the computational complexity of large road networks.

3. Preliminaries

3.1. Notations. Given an undirected graph $G = (V, E, X)$, where $V = \{v_1, v_2, \dots, v_n\}$ is a set of nodes with $|V| = n$, $E \subseteq V \times V$ is a set of edges that can be represented as an adjacency matrix $A \in \{0, 1\}^{n \times n}$, and $X = [x_1, x_2, \dots, x_n]^T \in \mathbb{R}^{n \times d_{in}}$ is a feature matrix with x_i denoting a feature vector of node $v_i \in V$. d_{in} is the length of the historical time series, and each feature in x_i corresponds to the traffic flow at a certain time. Our target is to obtain the traffic information $Y = [y_1, y_2, \dots, y_n]^T \in \mathbb{R}^{n \times d_{out}}$ (d_{out} is the length of traffic flow time series to be predicted) of a certain period of time in the future according to the historical traffic information X .

3.2. Graph Convolution Networks. As a semisupervised model, GCN can learn the hidden representation of each node. The hidden vectors of all nodes in layer $l + 1$ can be represented recursively by the hidden vectors of layer l as follows:

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-(1/2)} \tilde{A} \tilde{D}^{-(1/2)} H^{(l)} W^{(l)}\right), \quad (1)$$

where $\tilde{A} = A + I_n$, $W^{(l)}$ denotes the learnable weight matrix at layer l , $\tilde{D}_i = \sum_j \tilde{A}_{ij}$, and $\sigma(\cdot)$ is an activation function, such as ReLU. Initially, $H^{(0)} = X$.

4. Fast Graph Convolution Recurrent Neural Network

The traffic flow of a road is affected by the traffic flow of the surrounding roads and the historical traffic flow of the road itself, so the prediction model should consider these two factors. To model the temporal dependency of historical traffic on the road, GRU unit is embedded in the Seq2Seq model based on Encoder-Decoder framework to complete sequence prediction, and, to model the spatial correlation among neighbor roads, FastGCN is used in the traffic map of the road network to reduce the computational complexity and improve the efficiency. We integrate a model for quickly extracting spatiotemporal features, so we propose the

FastGCRNN (Fast Graph Revolution Recurrent Neural Network) model. The overall architecture of the model is shown in Figure 1.

This model mainly includes six parts, namely:

- (a) *Input Sequence X*. It is the input data of the whole prediction model, which is fed into the encoder part. In the road network traffic graph, it is the traffic flow of each node in a continuous period of time.
- (b) *Output Sequence Y*. It is the output of the whole prediction model (the output of decoder part). In the road network traffic graph, it is the traffic flow of each node road in the future.
- (c) *FastGCN Unit*. It can extract the spatial structure information of the road network through graph convolution. Based on that, it further uses sampling to reduce computational complexity.
- (d) *GRU Unit*. Traffic flows are time series signals, so we use GRU units to capture the long-term or short-term temporal dependence between the input traffic flow time series and embed two FastGCN units in its internal.
- (e) *Encoder Unit*. It is composed of GRU unit, and the output state of hidden layer is obtained by encoding the time series of the input traffic flow network graph.
- (f) *Decoder Unit*. It is also composed of GRU units. When it receives the encoder output, the decoder will continuously predict the traffic flow of each node.

The whole FastGCRNN model adopts the Seq2Seq model based on the Encoder-Decoder framework, which can use traffic flow of each road within the road network to predict the future traffic flow. Firstly, the continuous traffic flow data X on the road network is fed into the encoder part, and the data instance at each timestamp needs to go through FastGCN units to extract the spatial structure information between the road nodes, and then it needs to be processed by the GRU units in the encoder to get the temporal features of the traffic flow. After encoding, the hidden state output by the GRU units in the encoder is fed to the GRU units in decoder, and spatial features are further extracted by FastGCN. The final GRU units continuously predict the traffic flow Y .

4.1. Fast Spatial Feature Extractor: FastGCN. Each road in the urban road network does not exist in isolation but connects with the surrounding roads to form a whole. The traffic flow between roads is interactive; particularly, on the two-way road, there are vehicles flowing in and out. To model spatial correlation of traffic flows among road networks, we abstract the roads in road networks as nodes and their intersections as edges, as shown in Figure 2, where blue lines and dots represent road and intersections in road networks, respectively. Because we intend to predict traffic flows of the roads, while GCN can only make prediction on nodes, we model roads as nodes and their intersections as

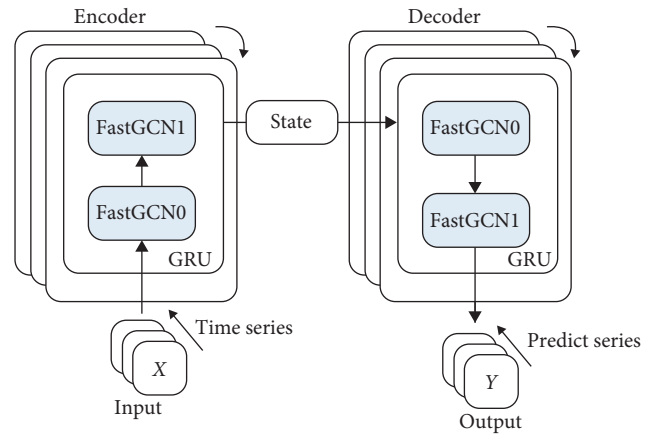


FIGURE 1: FastGCRNN model.

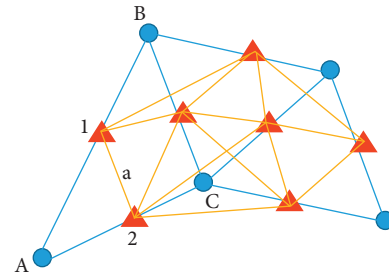


FIGURE 2: Construction process of road network graph.

edges, as illustrated through the red triangles and yellow lines in Figure 2, respectively.

In order to consider the influence of multihop in GCN, the number of layers of GCN will be increased recursively to realize the information exchange between multiple upstream and downstream roads. However, the recursive neighborhood expansion across layers poses time and memory challenges for training with large, dense graphs. To solve this problem, the FastGCN method is used, which interprets GCN as the integral transformation of the embedded function under the probability measure. The integration at this time can use the Monte Carlo method for consistency estimation, and the node training in the graph can also be performed in batches. Because the node training is carried out in batches, the structure of the graph is not limited; that is, when performing test prediction, the number of nodes and the connection relationship in the graph can change, and it does not have to be the same as the graph structure during training. This increases the generalization ability and scalability of the model to a certain extent.

The nodes in the graph of FastGCN can be regarded as independent and identically distributed sampling points that satisfy a certain probability distribution, and the calculated loss and convolution results are expressed as the integral form of the embedding function of each node. The estimation of integration can be expressed by Monte Carlo approximation which defines the sampling loss and sampling gradient. In order to reduce the variance of estimation, the sampling distribution can be further changed to make it more consistent with the real distribution. For example, the

simplest way is to use uniform distribution for sampling convolution. The improved method is to use importance sampling to make it continuously approach the real distribution and reduce the error caused by sampling.

If a node v in the graph G is taken as the observation object, its convolution can be considered as the information embedding expression of node v and all nodes in the graph in the upper layer through the addition of other forms of adjacency matrix and then the transformation of feature dimension through the trainable parameter matrix, which is equivalent to a discrete integral, and the adjacency matrix is equivalent to a weight given to each node. Therefore, the convolution process of node v in the graph is expressed in integral form as

$$\begin{aligned} \tilde{h}^{(l+1)}(v) &= \int \tilde{A}(v, u) h^{(l)}(u) W^{(l)} dP(u), \\ h^{(l+1)}(v) &= \sigma(\tilde{h}^{(l+1)}(v)), \quad l = 0, \dots, M-1. \end{aligned} \quad (2)$$

GCN in the form of integration is integrated by Monte Carlo method, and then it is transformed into the discrete form of sampling. At layer l , t_l points $(u_1^{(l)}, \dots, u_{t_l}^{(l)})$ are sampled independently and identically with probability p , and the approximate estimation is

$$\begin{aligned} \tilde{h}_{t_{l+1}}^{(l+1)}(v) &:= \frac{1}{t_l} \sum_{j=1}^{t_l} \tilde{A}(v, u_j^{(l)}) h_{t_l}^{(l)}(u_j^{(l)}) W^{(l)}, \\ \tilde{h}_{t_{l+1}}^{(l+1)}(v) &:= \sigma(\tilde{h}_{t_{l+1}}^{(l+1)}(v)), \quad l = 0, \dots, M-1. \end{aligned} \quad (3)$$

If each layer of convolution uses this method for sampling and information transfer, after layer M , the embedded expression of node v is

$$\begin{aligned} H^{(l+1)}(v, :) &= \sigma\left(\frac{n}{t_l} \sum_{j=1}^{t_l} \tilde{A}(v, u_j^{(l)}) H^{(l)}(u_j^{(l)}, :) W^{(l)}\right), \\ & \quad l = 0, \dots, M-1. \end{aligned} \quad (4)$$

In the above integral form of GCN, the embedded information expression of node V needs to be obtained from all nodes in the graph. However, after sampling, only t_l

nodes in the graph need to exchange and fuse information in FastGCN, so the calculation complexity of the whole graph changes from n^2 to $(t_l \times n)$, and the efficiency is greatly improved.

Here is an example to illustrate the advantages of FastGCN compared with GCN, if the abstract road network graph has 5 nodes and 6 edges, as shown in Figures 3 and 4.

In GCN, each epoch must be put into a complete graph, instead of using only a few nodes in the graph; that is, each node in the graph needs to convolute and exchange information with all other nodes in the graph. In FastGCN, we decompose the large graph into several small graphs by batch operation and put them into memory, as well as the method of sampling to remove the information exchange with some low correlation nodes. Each node only interacts with the sampled nodes in the graph. As shown in Figure 4, each node only interacts with node A and node E. In this way, the computing efficiency is greatly improved, especially when it can be calculated on a large graph without memory overflow.

For the sampling method, in order to make the sampling closer to the real connected nodes, FastGCN does not use uniform sampling [40], but importance sampling. That is, each node is not sampled according to the same probability, but using probability distribution Q . No matter what probability distribution sampling is used, the mean value of the sample is constant, but it will affect the variance of the sample. In order to minimize the error, the distribution Q which can minimize the sample variance is selected here. At this time, the calculation output of node v passing through FastGCN layer is

$$\begin{aligned} H^{(l+1)}(v, :) &= \sigma\left(\frac{1}{t_l} \sum_{j=1}^{t_l} \frac{\tilde{A}(v, u_j^{(l)}) H^{(l)}(u_j^{(l)}, :) W^{(l)}}{q(u_j^{(l)})}\right), \\ & \quad u_j^{(l)} \sim q, \quad l = 0, \dots, M-1. \end{aligned} \quad (5)$$

In the experiment, only two FastGCN units were used to extract spatial features. This is because we need to avoid the problem of oversmoothing [41]. The specific calculation process is as follows:

$$f(\tilde{A}, X) = \sigma\left(\frac{1}{t_l} \sum_{j=1}^{t_l} \frac{\tilde{A}(v, u_j^{(1)}) \sigma\left(\frac{1}{t_l} \sum_{j=1}^{t_l} \left(\frac{\tilde{A}(v, u_j^{(0)}) X(u_j^{(0)}, :) W^{(0)}}{q(u_j^{(0)})}\right) (u_j^{(1)}, :) W^{(1)}\right)}{q(u_j^{(1)})}\right), \quad u_j^{(0)} \sim q, u_j^{(1)} \sim q. \quad (6)$$

4.2. Fast Temporal Feature Extractor: GRU. This is a key issue to effectively capture the long-term temporal dependence of traffic flow. The observed value of each timestamp is shown in Figure 5. The flow value of each node will change with time. The prediction is a typical time series prediction problem; that is, given the observed value of each road at d_{in} timestamps in history, the traffic flow value of d_{out} timestamps in the future will be predicted.

LSTM and GRU are commonly used in time series prediction. Both models use gating mechanisms to remember as much long-term information as possible and are equally effective for various tasks. To maximize efficiency, we chose GRU with relatively simple structure, fewer parameters, and faster training ability. GRU unit has update gate, reset gate, and memory unit, which can make it have a process of screening memory for historical data, so it can

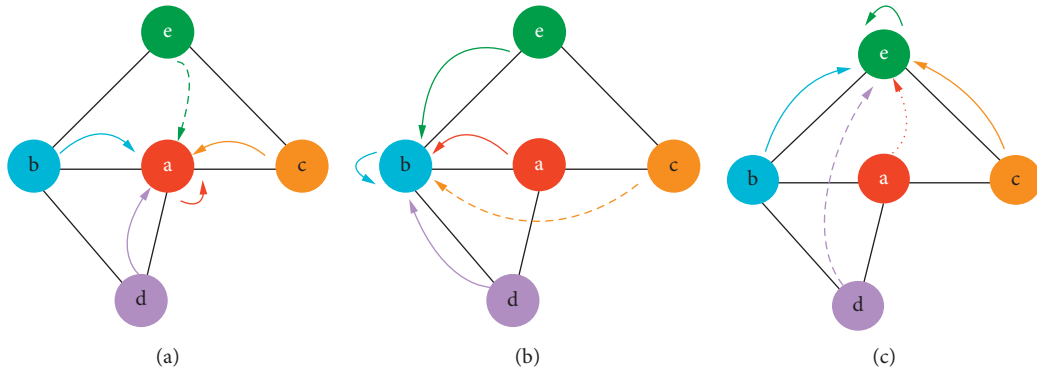


FIGURE 3: The process of GCN performing a convolution operation. (a) Convolution process of node A. (b) Convolution process of node B. (c) Convolution process of node E.

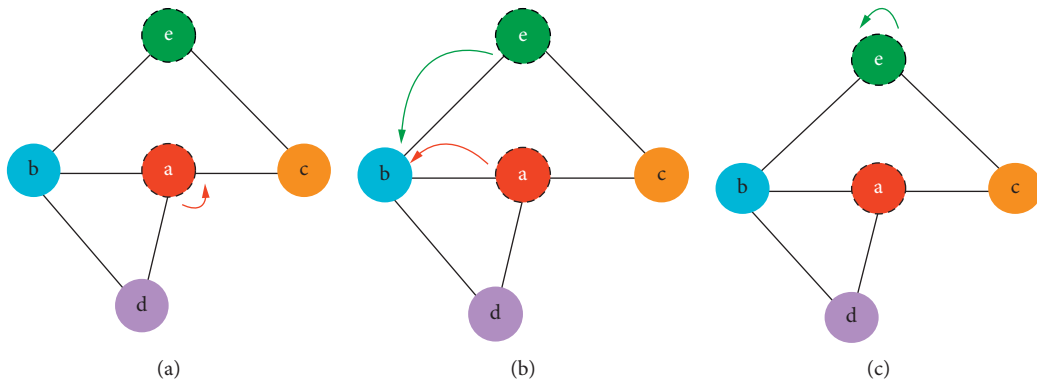


FIGURE 4: Convolution operation process in a batch of FastGCN under sampling distribution. (a) Sampling convolution operation of node A. (b) Sampling convolution operation of node B. (c) Sampling convolution operation of node E.

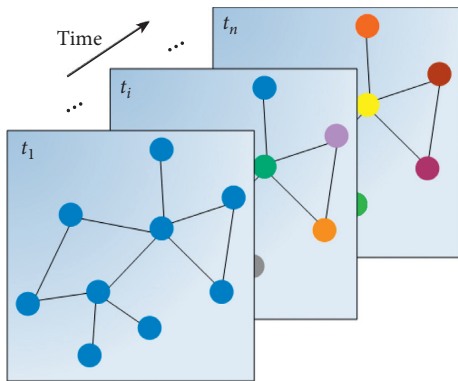


FIGURE 5: Traffic flow data with graph structure at different timestamps.

retain long-term memory. In GRU, time sequence information is saved by memory unit, which can capture long- and short-term memory in time and improve the accuracy of prediction.

In order to complete the sequence prediction, the Seq2Seq model based on the Encoder-Decoder structure is used. Seq2Seq puts the input history sequence into GRU, extracts the timing features, and obtains the hidden state vector C of the input sequence as the coding result of the encoder. This state vector C contains the feature information

of all previous moments, which is a centralized embodiment of their temporal features. In the decoder, C is used as the initial input of decoder to generate the predicted time series. In this way, Seq2Seq can extract the temporal characteristics of the traffic volume in the previous period, such as the proximity, trend, and periodicity of the traffic flow in the time dimension. When predicting the traffic volume, the model can obtain a smoothly changing traffic volume according to the proximity, and the characteristics of the proximity can be adjusted according to the trend and periodicity.

5. Experiment

In order to illustrate the role of the model in the large graph, 1865 roads in Luohu District of Shenzhen city are selected for the experiment, and the specific roads and areas are shown in Figure 6.

To calculate the traffic flows in each road, we map the GPS coordinates to the corresponding roads through the Frechet method [42]. The format of the mapped data is shown in Table 1. The core fields are road number (road_id), license plate number (car_id), and upload time (time). Each data record represents the information; the taxis with the car_id are on the road with road_id at the specific time.

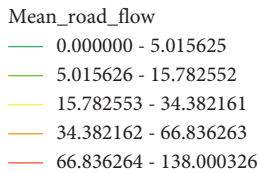


FIGURE 6: Part of the road network map of Luohu District, Shenzhen.

TABLE 1: Shenzhen taxi GPS record information example.

Road_id	Car_id	Time
92230	02341	2015-01-01 00:03:46
92230	03982	2015-01-02 06:23:12
...

5.1. Data Preprocessing. In data preprocessing, the taxi data in Shenzhen is transformed into the form of continuous timestamps on the road network, i.e., the traffic data shown in Figure 5. Specifically, we map the original GPS upload data to the road and count the traffic flow on each road in each time period. The data preprocessing algorithm is shown in Algorithm 1.

5.2. Comparative Experiment. The biggest advantage of FastGCRNN model is that it can be applied to large graphs, and it can reduce the computational complexity without losing the accuracy of the model. On the road network data of Shenzhen, the experiment is conducted with the traffic flow series of different time intervals to compare with some classic traffic flow prediction models: (1) HA, (2) ARIMA, (3) SVR, (4) LSTM, (5) ConvLSTM, (6) GCRN [18], and (7) GCRNN-nosample. The evaluation standard used in the experiment is Root Mean Squared Error (RMSE) [43]. The specific experimental results are shown in Table 2.

From the table results, we can find that FastGCRNN model has reached the best prediction performance in terms of RMSE. In these comparison models, HA, ARIMA, SVR, and LSTM only consider the temporal correlation without considering the spatial correlation, which is also one of the reasons for their poor accuracy. ConvLSTM divides the urban area into a grid and maps the traffic volume in each time period to the grid, and the traffic volume is regarded as the pixel value of the grid. Although this method considers

- (1) **Initialize:** $time_interval = 5$ min (or 30 min),
 $begin_time = 2015-01-01\ 00:00:00$,
 $roadflow[roadid][time_num] = 0$
- (2) **For Every data record do**
- (3) $time_num \leftarrow (time - begin_time) / time_interval$
- (4) **End for**
- (5) All data records are grouped by car_id , sorted by $time_num$ within the group
- (6) **For each group records do**
- (7) Remove duplicates records based on $road_id$ and $time_num$
- (8) Count $roadflow$
- (9) **End for**
- (10) **Output:** $roadflow$

ALGORITHM 1: Generate traffic flow time series for different roads.

TABLE 2: Comparison of results between FastGCRNN model and other traffic flow prediction models.

Model	Time	
	RMSE	
	5 min	30 min
HA	19.502	23.158
ARIMA	17.541	19.097
SVR	17.895	19.005
LSTM	13.102	16.930
ConvLSTM	19.481	21.038
GCRN	11.892	16.265
GCRNN-nosample	9.950	16.231
FastGCRNN	9.867	16.2734

the spatial correlation of vehicle flow, it also loses the topological structure relationship of the road network graph.

For verification, the proposed GCRNN can reduce the computational complexity, compared with the GCRN model, which also captures the topology information of the road network; the result is shown in Figure 7.

In Figure 7, we only compare the baselines with higher prediction accuracy, namely, GCRN and GCRNN-nosample. From Figure 7, it can be observed that the computational complexity of FastGCRNN is the lowest. The training time of FastGCRNN is about 0.03 times that of GCRN. Moreover, FastGCRNN reduces the training time to 1/3 times that of GCRNN-nosample, i.e., the GCRNN model without sampling. From the experiment results, it can be concluded that both the GCRNN model and the sampling method can reduce the training time.

5.3. Model Parameter Analysis. In FastGCRNN, each sampling point has a certain effect on the accuracy and training time of the model. When using 1685 roads in Shenzhen for experiments, different sampling sizes were set to compare the accuracy and time changes. The experimental results are shown in Figure 8. The abscissa in the figure shows the sampling size of FastGCN unit in the first and second layers, respectively. The blue column represents the RMSE of the

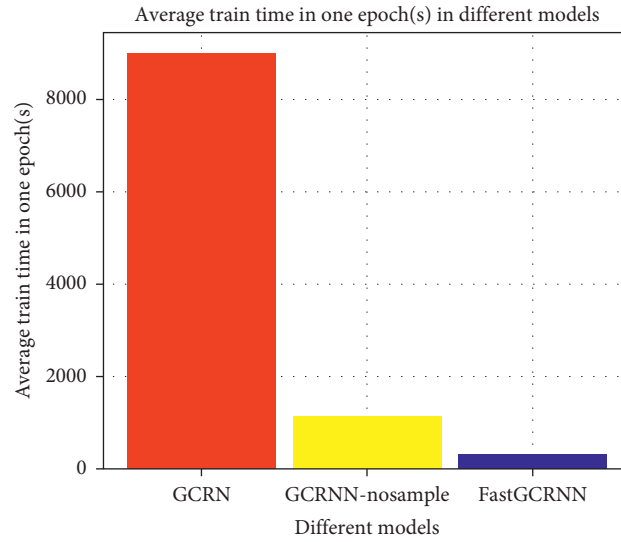


FIGURE 7: Time consumption of training an epoch with different models.

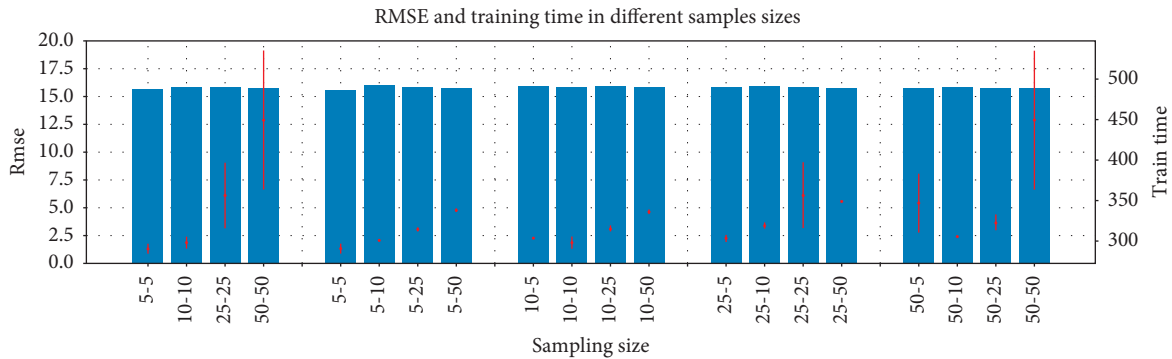


FIGURE 8: RMSE and training time when using different sampling sizes in two layers of FastGCN.

prediction results. The red line indicates the time consumption in each epoch, and the upper and lower ends are the maximum and minimum values of time consumption in the training process.

From the experimental results, it can be seen that choosing different sampling sizes has little effect on accuracy, and it does not necessarily mean that the more the samples, the more the information obtained, and the better the prediction effect. For example, the accuracy of sampling 50 nodes for each layer in the figure is not the best, because there are “bridge” type (other nodes affecting the central node will spread to other unrelated distant areas) and “tree” type (other nodes affecting the central node will be limited to the small area to which the node belongs) of connection relationship between nodes [44]. If more nodes are sampled, the influence relationship of the nodes will spread to unrelated areas, resulting in information redundancy, misleading the update of node features, and reducing the prediction accuracy. In addition, in the road network graph, intersections generally connect four roads; that is to say, selecting four nodes in one hop can complete the extraction of feature information. Here is the statistics of 1865 selected roads’ degrees, as shown in Figure 9. Among them, the nodes

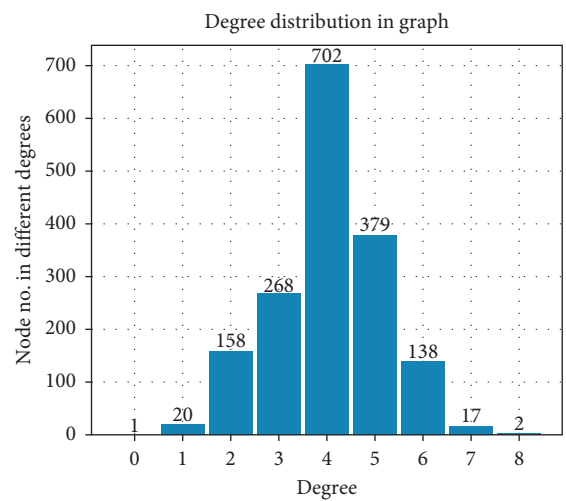


FIGURE 9: Distribution of node degree of road network graph in Shenzhen.

with degree 4 are the most, and the degrees of 70% of the nodes are less than 5, and the degrees of nearly 99% of the nodes are less than 7. Therefore, the case of sampling size 5

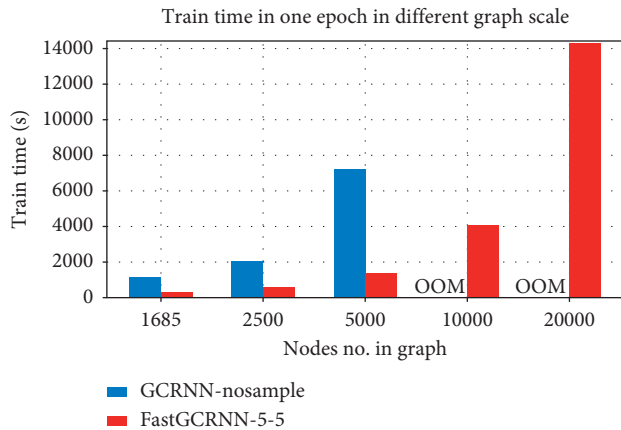


FIGURE 10: Time consumption of FastGCRNN and GCRNN unsampled models at different graph sizes.

can already include the neighbors in one hop around it. In this case, not only the training time is reduced, but also the accuracy is not reduced.

And we compared the time consumption of FastGCN and standard GCN in different sizes of graphs. The experimental results are shown in Figure 10.

From the experimental results, it can be seen that FastGCRNN has obvious advantages in dealing with large graph problems. Particularly, when the size of graph reaches a certain degree, FastGCRNN is still running normally when GCRNN-nosample model has overflowed memory and cannot be trained.

6. Conclusions

This paper mainly deals with the problem of large graphs with spatiotemporal properties by constructing the FastGCRNN model and applies them to road network traffic graphs. The model predicts the traffic flow by extracting the temporal and spatial attributes of the traffic flow on the large-scale road networks. Among them, FastGCN is used to extract the topological structure in the space and accelerate training and reduce complexity. GRU is used to extract time series features, and the Seq2Seq model based on the Encoder-Decoder framework can complete sequence prediction tasks of unequal length. The most prominent advantage of this model is the FastGCN embedded in it, which uses the sampling method to accelerate the extraction of spatial features, reduce computational complexity, and improve efficiency. Moreover, the model is not prone to memory overflow in processing large-scale graph-structured data.

It is worth mentioning that this model is not only applicable to traffic flow data, but also applicable to all graph structure data with spatiotemporal characteristics, especially the largerscale data.

Data Availability

The data used to support the findings of this study are available upon request to Ya Zhang, zndxxxxzy@csu.edu.cn.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

- [1] U. Mori, A. Mendiburu, M. Álvarez, and J. A. Lozano, "A review of travel time estimation and forecasting for advanced traveller information systems," *Transportmetrica A: Transport Science*, vol. 11, no. 2, pp. 119–157, 2015.
- [2] Y. Zhang, T. Cheng, and Y. Ren, "A graph deep learning method for short-term traffic forecasting on large road networks," *Computer-Aided Civil and Infrastructure Engineering*, vol. 34, no. 10, pp. 877–896, 2019.
- [3] P. Wang, J. Lai, Z. Huang, Q. Tan, and T. Lin, "Estimating traffic flow in large road networks based on multi-source traffic data," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–12, 2020.
- [4] A. Reggiani and L. A. Schintler, *Introduction: Cross Atlantic Perspectives in Methods and Models Analysing Transport and Telecommunications*, Springer Science & Business Media, Berlin, Germany, 2005.
- [5] M. S. Ahmed and A. R. Cook, *Analysis of Freeway Traffic Time-Series Data by Using Box-Jenkins Techniques*, Springer, Berlin, Germany, 1979.
- [6] G. A. Davis and N. L. Nihan, "Nonparametric regression and short-term freeway traffic forecasting," *Journal of Transportation Engineering*, vol. 117, no. 2, pp. 178–188, 1991.
- [7] I. Okutani and Y. J. Stephanedes, "Dynamic prediction of traffic volume through Kalman filtering theory," *Transportation Research Part B: Methodological*, vol. 18, no. 1, pp. 1–11, 1984.
- [8] C. Kim and A. G. Hobeika, "Short-term demand forecasting model from real-time traffic data," in *Proceedings of the Infrastructure Planning and Management*, pp. 540–550, Denver, CO, USA, June 1993.
- [9] X. Luo, L. Niu, and S. Zhang, "An algorithm for traffic flow prediction based on improved SARIMA and GA," *KSCCE Journal of Civil Engineering*, vol. 22, no. 10, pp. 4107–4115, 2018.
- [10] N. K. Chikkakrishna, C. Hardik, K. Deepika, and N. Sparsha, "Short-term traffic prediction using sarima and FbPROPHET," in *Proceedings of the 2019 IEEE 16th India Council International Conference, INDICON 2019-Symposium Proceedings*, pp. 1–4, Rajkot, Gujarat, India, December 2019.
- [11] B. Liu, X. Tang, J. Cheng, and P. Shi, "Traffic flow combination forecasting method based on improved LSTM and ARIMA," *International Journal of Embedded Systems*, vol. 12, no. 1, pp. 22–30, 2020.
- [12] B. Yang, S. Sun, J. Li, X. Lin, and Y. Tian, "Traffic flow prediction using LSTM with feature enhancement," *Neurocomputing*, vol. 332, pp. 320–327, 2019.
- [13] G. Dai, C. Ma, and X. Xu, "Short-term traffic flow prediction method for urban road sections based on space-time analysis and GRU," *IEEE Access*, vol. 7, pp. 143025–143035, 2019.
- [14] P. Li, M. Sun, and M. Pang, "Prediction of taxi demand based on convLSTM neural network," in *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11305, pp. 15–25, Springer, Berlin, Germany, 2018.
- [15] R. He, N. Xiong, L. T. Yang, and J. H. Park, "Using multi-modal semantic association rules to fuse keywords and visual features automatically for web image retrieval," *Information Fusion*, vol. 12, no. 3, pp. 223–230, 2011.

- [16] X. Shi, Z. Chen, H. Wang, D. Y. Yeung, W. K. Wong, and W. C. Woo, "Convolutional LSTM network: a machine learning approach for precipitation nowcasting," in *Advances in Neural Information Processing Systems*, pp. 802–810, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2015.
- [17] Y. Li, R. Yu, C. Shahabi, and Y. Liu, "Diffusion convolutional recurrent neural network: data-driven traffic forecasting," 2017, <http://arxiv.org/abs/1707.01926>.
- [18] Y. Seo, M. Defferrard, P. Vandergheynst, and X. Bresson, "Structured sequence modeling with graph convolutional recurrent networks," in *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11301, pp. 362–373, Springer, Berlin, Germany, 2018.
- [19] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, "Spectral networks and locally connected networks on graphs," 2013, <http://arxiv.org/abs/1312.6203>.
- [20] C. Lin, Y.-X. He, and N. Xiong, "An energy-efficient dynamic power management in wireless sensor networks," in *Proceedings of the 2006 Fifth International Symposium on Parallel and Distributed Computing*, pp. 148–154, Rhodes Island, Greece, April 2006.
- [21] Y. Liu, M. Ma, X. Liu, N. Xiong, A. Liu, and Y. Zhu, "Design and analysis of probing route to defense sink-hole attacks for internet of things security," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 1, 2018.
- [22] L. Shu, Y. Zhang, Z. Yu, L. T. Yang, M. Hauswirth, and N. Xiong, "Context-aware cross-layer optimized video streaming in wireless multimedia sensor networks," *The Journal of Supercomputing*, vol. 54, no. 1, pp. 94–121, 2010.
- [23] Y. Wang, A. V. Vasilakos, J. Ma, and N. Xiong, "On studying the impact of uncertainty on behavior diffusion in social networks," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 45, no. 2, pp. 185–197, 2014.
- [24] H. Zheng, W. Guo, and N. Xiong, "A kernel-based compressive sensing approach for mobile data gathering in wireless sensor network systems," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 48, no. 12, pp. 2315–2327, 2017.
- [25] Z. Wan, N. Xiong, N. Ghani, A. V. Vasilakos, and L. Zhou, "Adaptive unequal protection for wireless video transmission over IEEE 802.11e networks," *Multimedia Tools and Applications*, vol. 72, no. 1, pp. 541–571, 2014.
- [26] J. Li, N. Xiong, J. H. Park, C. Liu, S. Ma, and S. Cho, "Intelligent model design of cluster supply chain with horizontal cooperation," *Journal of Intelligent Manufacturing*, vol. 23, no. 4, pp. 917–931, 2012.
- [27] W.-L. Chiang, X. Liu, S. Si, Y. Li, S. Bengio, and C.-J. Hsieh, "Cluster-gcn: an efficient algorithm for training deep and large graph convolutional networks," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 257–266, Anchorage, AK, USA, June 2020.
- [28] J. Chen, T. Ma, and C. Xiao, "FastGCN: Fast learning with graph convolutional networks via importance sampling," in *Proceedings of the 6th International Conference on Learning Representations, ICLR 2018-Conference Track Proceedings*, Vancouver, Canada, May 2018.
- [29] Z. Wang, T. Li, N. Xiong, and Y. Pan, "A novel dynamic network data replication scheme based on historical access record and proactive deletion," *The Journal of Supercomputing*, vol. 62, no. 1, pp. 227–250, 2012.
- [30] Y. Yang, N. Xiong, N. Y. Chong, and X. Défago, "A decentralized and adaptive flocking algorithm for autonomous mobile robots," in *Proceedings of the 2008 The 3rd International Conference on Grid and Pervasive Computing-Workshops*, pp. 262–268, Kunming, China, May 2008.
- [31] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems*, pp. 3104–3112, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2015.
- [32] Z. Pan, Y. Liang, W. Wang, Y. Yu, Y. Zheng, and J. Zhang, "Urban traffic prediction from spatio-temporal data using deep meta learning," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1720–1730, Anchorage, AK, USA, July 2019.
- [33] Q. Hou, J. Leng, G. Ma, W. Liu, and Y. Cheng, "An adaptive hybrid model for short-term urban traffic flow prediction," *Physica A: Statistical Mechanics and Its Applications*, vol. 527, Article ID 121065, 2019.
- [34] Y. Zeng, C. J. Sreenan, N. Xiong, L. T. Yang, and J. H. Park, "Connectivity and coverage maintenance in wireless sensor networks," *The Journal of Supercomputing*, vol. 52, no. 1, pp. 23–46, 2010.
- [35] C. Lin, N. Xiong, J. H. Park, and T.-h. Kim, "Dynamic power management in new architecture of wireless sensor networks," *International Journal of Communication Systems*, vol. 22, no. 6, pp. 671–693, 2009.
- [36] Y. Sang, H. Shen, Y. Tan, and N. Xiong, "Efficient protocols for privacy preserving matching against distributed datasets, information and communications security," in *Proceedings of the International Conference on Information and Communications Security*, pp. 210–227, Raleigh, NC, USA, December 2006.
- [37] F. Long, N. Xiong, A. V. Vasilakos, L. T. Yang, and F. Sun, "A sustainable heuristic QoS routing algorithm for pervasive multi-layered satellite wireless networks," *Wireless Networks*, vol. 16, no. 6, pp. 1657–1673, 2010.
- [38] W. Guo, N. Xiong, A. V. Vasilakos, G. Chen, and C. Yu, "Distributed k-connected fault-tolerant topology control algorithms with PSO in future autonomic sensor systems," *International Journal of Sensor Networks*, vol. 12, no. 1, pp. 53–62, 2012.
- [39] N. Xiong et al., "A self-tuning failure detection scheme for cloud computing service," in *Proceedings of the 2012 IEEE 26th International Parallel and Distributed Processing Symposium*, pp. 668–679, Shanghai, China, May 2012.
- [40] M. Setia, "Methodology series module 5: sampling strategies," *Indian Journal of Dermatology*, vol. 61, no. 5, p. 505, 2016.
- [41] Q. Li, Z. Han, and X. M. Wu, "Deeper insights into graph convolutional networks for semi-supervised learning," in *Proceedings of the 32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pp. 3538–3545, New Orleans, LA, USA, February 2018.
- [42] X. Song, V. Raghavan, and D. Yoshida, "Matching of vehicle GPS traces with urban road networks," *Current Science*, vol. 98, no. 12, pp. 1592–1598, 2010.
- [43] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)?-arguments against avoiding RMSE in the literature," *Geoscientific Model Development*, vol. 7, no. 3, pp. 1247–1250, 2014.
- [44] K. Xu, C. Li, Y. Tian, T. Sonobe, K. I. Kawarabayashi, and S. Jegelka, "Representation learning on graphs with jumping knowledge networks," in *Proceedings of the 35th International Conference Machine Learning ICML 2018*, pp. 8676–8685, Stockholm Sweden, July 2018.

Research Article

Research on Coordinated Development of a Railway Freight Collection and Distribution System Based on an “Entropy-TOPSIS Coupling Development Degree Model” Integrated with Machine Learning

Yun Jing, Si-Ye Guo , Xuan Wang , and Fang-Qiu Chen 

School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China

Correspondence should be addressed to Si-Ye Guo; 18120799@bjtu.edu.cn

Received 4 June 2020; Revised 24 July 2020; Accepted 10 August 2020; Published 15 September 2020

Academic Editor: Naixue Xiong

Copyright © 2020 Yun Jing et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In recent years, with the gradual networking of high-speed railways in China, the existing railway transportation capacity has been released. In order to improve transportation capacity, railway freight transportation enterprises companies have gradually shifted the transportation of goods from dedicated freight lines to passenger-cargo lines. In terms of the organization form of collection and distribution, China has a complete research system for heavy-haul railway collection and distribution, but the research on the integration of collection and distribution of the ordinary-speed railway freight has not been completed. This paper combines the theories of the integration of collection and distribution theory, coordination theory, and coupling theory and incorporates the machine learning fuzzy mathematics to construct an “Entropy-TOPSIS Coupling Development Degree Model” for dynamic intelligent quantitative analysis of the synergy of railway freight collection and distribution systems. Finally, we take the Tongchuan Depot of “China Railway Xi’an Group Co., Ltd.” as a research object to construct a target system and use the intelligent information acquisition system to collect basic data. The analysis results show that through the coordinated control of the freight collection and distribution system, the coordination between the subsystems of the integrated freight collection and distribution system is increased by 5.94%, which verifies the feasibility of the model in the quantitative improvement of the integration of collection and distribution system. It provides a new method for the research of integrated development of railway freight collection and distribution.

1. Introduction

In recent years, China’s high-speed railway construction has been developing very rapidly. By the end of 2019, the operating mileage of high-speed railways has exceeded 35,000 kilometers, leading to a widely-covered high-speed railway network between major cities. With the rapid advancement of the high-speed railway network, the existing railway transportation capacity has been released, which makes it possible for railway transportation enterprises to gradually expand the freight organization from the dedicated freight line to the passenger-cargo line to improve freight transportation capacity. With the rapid development of various modes of transportation, the

transportation mode is moving towards the direction of collaboration, cooperation and the establishment of joint transportation system. To meet the needs of railway logistics and to improve railway freight volume and the transportation network’s efficiency, it is of vital importance to integrate the freight collection and distribution based on railway transportation channels and accelerate the speed of cargo assembly and delivery. However, the traditional static traffic flow control is no longer suitable for the current dynamic changing collection and distribution system. On the other hand, the integrated transportation organization, as well as dynamic control and integration of the railway freight collection and distribution system, will both help to coordinate the overall

integrated configuration, which plays a key role in the coordinated development of railway collection and distribution system, such as improving the planning and layout, coordinating the point and line capacity, and matching of technical standards. Simultaneously, the research is conducive to actively carrying out collaborative marketing, realizing dynamic coordination and distribution of empty and heavy cargoes, expending the transportation market share of transportation, and achieving the purpose of unblocked transportation and maximum benefits.

This paper provides a new method for the research of integrated development of railway freight collection and distribution system. Based on the theories of integration of collection and distribution theory, synergy theory, coupling theory, and fuzzy mathematics, we construct an “Entropy-TOPSIS Coupling Development Degree Model” for dynamic and intelligent quantitative analysis of the degree of synergy between collection and distribution systems. Through the Entropy-TOPSIS model, we obtain the collaborative development evaluation index, which can help us make a longitudinal evaluation of the comprehensive coordination situation of the railway freight collection and distribution system. On this basis, we also construct the coupling development degree model to obtain the coupling development degree among the various subsystems, so that we can quantify the synergy between the collection and distribution integrated systems. The result provides a substantive reference basis for railway transportation companies to further strengthen the integrated regulation of freight collection and distribution system, which also plays an important role in the planning and construction of the railway collection and distribution system channel, the optimization of resource allocation, the improvement of transportation capacity and the diversified development of railway freight transportation.

The remainder of this study is organized as follows: Section 2 summarizes the related literature on the collection and distribution system. Section 3 clarifies the cooperative connotation of the integration of the railway freight collection and distribution system, and establishes an evaluation index system for the cooperative configuration of the railway freight collection and distribution system. Section 4 constructs a mathematical model, called the “Entropy-TOPSIS Coupling Development Degree Model” for quantitatively analyzing the synergy of the railway freight collection and distribution system. Section 5 takes Tongchuan Depot, which is a sub-division of “China Railway Xi’an Group Co., Ltd.”, as a research object and carries out a quantitative analysis of the coordinated development of the target collection and distribution system, to verify the feasibility of the model. Finally, Section 6 offers conclusions and future research directions.

2. Literature Review

The concept of the collection and distribution system theory first appeared in international port cargo transportation. Having a suitable geographical location and an efficient collection and distribution network is a guarantee for the

port healthy development. With the development of logistics transportation, a comprehensive collection and distribution system becomes an effective way to diversify the port. Article [1, 2] pointed out from the perspective of shipping links that the development of port collection and distribution system was mainly affected by the conditions of the collection and distribution infrastructure, the adaptability of the collection and distribution method, and the environment. Huang [3] constructed an optimization model of the Shanghai Port container transportation system for the port collection and distribution system and predicted the development bottleneck of the Shanghai Port container transportation collection and distribution system. Geng [4] put forward the basic characteristics of the medium and long-term development of the port collection and distribution railway and the problems that the current development needs to address, which provided theoretical support for promoting the development of combined transportation of railway and water, accelerated the construction of an integrated transportation system, and achieved cost reduction and efficiency increase in the logistics industry. Xu [5] analyzed the current situation and existing problems of the collection and distribution system of Tonghai Port Area, and provided reference for the connection of the special line of the Shugang Railway and the station setting. The research of the heavy-haul railway collection and distribution system is also relatively complete. Wu [6] analyzed the main coordination factors of the railway transportation system in the Caofeidian port area from three aspects: facilities, production organization and operation management and used cloud model to evaluate the coordination of the railway transportation system. Kong and He [7] explored the capacity coordination of the Shenhua heavy-duty coal transportation special line collection and distribution system from the three subsystems. Besides, they established a capacity coordination model for the collection and distribution system of the heavy-haul coal transportation dedicated line, and proposed a classification scheme for the coordination degree to provide a reference for optimizing the heavy-haul railway collection and distribution system. Feng et al. [8] analyzed the connotation and synergy motivation of the integrated organization of the railway heavy-duty collection and distribution system and believed that it was determined by the self-organization, instability principle, dominating principle, and order parameter principle of the railway heavy-duty transportation system. From the system perspective, Yu [9] analyzed the heavy-haul railway collection and distribution system and proposed research methods. The results show that the method can not only effectively meet the transportation target effectively, but realize the expansion and transformation of the railway transportation subsystems as well. Therefore, it can meet future development needs.

As for the freight flow control, the allocation of empty and heavy cargoes is an important part of the railway freight collection and distribution link. Therefore, the realization of the dynamic control of the flow of cargoes will facilitate the alleviation of the operating pressure of busy stations and the improvement of the efficiency of integrated operations. Liang and Lin [10] constructed a strategic optimization

model for the organization of dynamic train flow in railway transportation. The model combined heavy and empty cargo transportation and designed an improved genetic algorithm based on integer coding that can solve large-scale network problems. Regarding the bottleneck section of the railway network, Wang et al. [11] established a multiobjective planning model for the optimal allocation of transport capacity resources in the bottleneck section, and verified the rationality and effectiveness of the method for calculating traffic flow and the optimization of transport capacity resource allocation in the bottleneck section. Xue et al. [12] provided the calculation method of the coupling degree between the station stage plan and the dynamic traffic flow. The results of the calculation example showed that the proper traffic flow allocation can improve the coupling degree between the stage plan and the dynamic traffic flow to a certain extent.

In terms of the research theory, this paper mainly deals with collaborative marketing in synergy. Varadarajan [13] summarized the connotation and application scope of symbiotic marketing, studied the impact of business environment changes and organizational development on symbiotic marketing, and discussed the feasibility of symbiotic marketing and its planning and execution. Xie [14] started from the definition of collaborative marketing, analyzed the principles of selecting collaborative objects, and gave suggestions on how to choose collaborative objects for enterprises. As for the research methods, this paper mainly involves evaluation cooperative model and coordinated development model. Both are often used in the fields of economics and systems science. Sun [15] used fuzzy comprehensive evaluation and a method with preference to evaluate partners. Yin and Bao [16] used the Entropy-TOPSIS method to conduct financial risk evaluation research on high-tech enterprises. Wang and Tang [17], taking the Dongting Lake area as a research object, and based on the construction of an evaluation index system for the coordinated development of the Ecological-Economic-Society complex system, introduced the coupling coordination degree function was introduced to empirically measure the coupling degree, coordination degree, and comprehensive development value, among the subsystems. In view of the coordination of regional development under the background of high-speed railway. Zhang et al. [18] established a High-speed railway-Population-Economy coupling coordination degree evaluation index system to measure the coupling coordination degree of 27 provinces across the country, and analyzed the spatial connections among them.

From the above literature review, the following points can be seen. To sum up, the following points can be seen. For the research object, the collection and distribution technology is widely used in the port and multimodal transportation. However, in the railway freight transportation, this research has a limitation on the single line or a single transportation organization mode. For freight flow scheduling, realizing the dynamic real-time control of train flow is an imperative measure for railway transportation enterprises; for the research method, the Entropy-TOPSIS model which method can objectively evaluate the target system, is often adopted in the railway field, while the coupling development degree

model, though it has been fully utilized in the economic and system science field, is rarely applied in the railway field. Hence, the combination of the Entropy-TOPSIS model and coupling development degree model into the integrated development of the collection and distribution system will generate some new discoveries.

3. Synergy Connotation and Characteristics of Integrated Railway Freight Collection and Distribution

3.1. The Main Connotation of the Integrated Railway Collection and Distribution Transportation System. As a system platform, the railway collection and distribution system connects the railway loading station, road trunk and branch line, hub and technical operation station, and unloading station, realizes the balanced transportation of goods between the production and consumption places.

The railway collection and distribution system is a complete integrated logistics chain, which is usually completed by the railway and other transportation modes [19]. The railway collection and distribution system is mainly composed of three subsystems: “collection system”, “transportation system”, and “digestion system”. The subsystems are interrelated and restricted with each other and together form a “link” for the flow of goods, cargoes and trains [19]. The flow diagram of the railway collection and distribution system is shown in Figure 1.

3.2. Synergy Connotation and Characteristics of the Integration of Railway Freight Collection and Distribution. Synergys is a theory founded established by the German physicist Haken in the 1970s to study how the various subsystems in a complex and complex system work together well. It is an important branch of the system science.

The connotation of the integrated integration of railway freight collection and distribution is: under the condition of a certain level of transportation organization, through the close cooperation between various units and departments within the railway enterprise and the cooperation of the “Consolidation-Dispersion” side, coordinate and, complement each other’s functions, and realize the capabilities of the “aggregation,” “dispersion,” and “transportation” subsystems are coordinated to achieve the overall optimum. Meanwhile, the integration of railway freight collection and distribution can actively carry out collaborative marketing, expand the transportation market share, and achieve the goal of unblocked transportation and maximum benefits.

3.3. Evaluation Index System. Based on the operation process of the railway freight collection and distribution system, we combine with the components of the collection and distribution system. Then, from the perspective of the three subsystems of consolidation, transportation, and dispersion system, we analyze, the railway freight collection and distribution system coordination evaluation index system, which is shown in Figure 2.

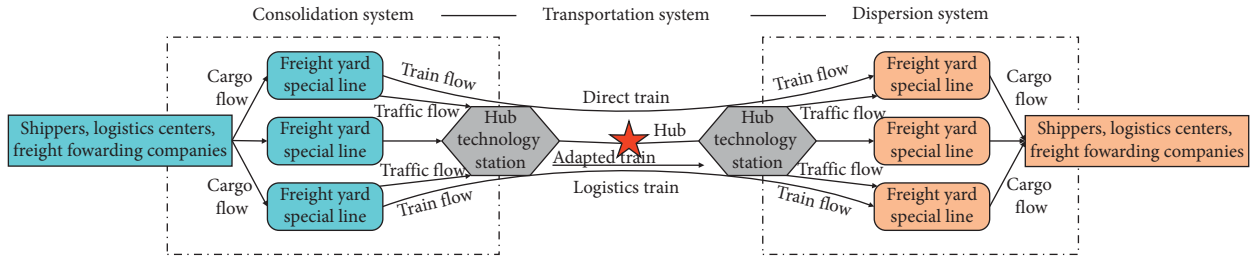


FIGURE 1: Schematic diagram of the railway collection and distribution system.

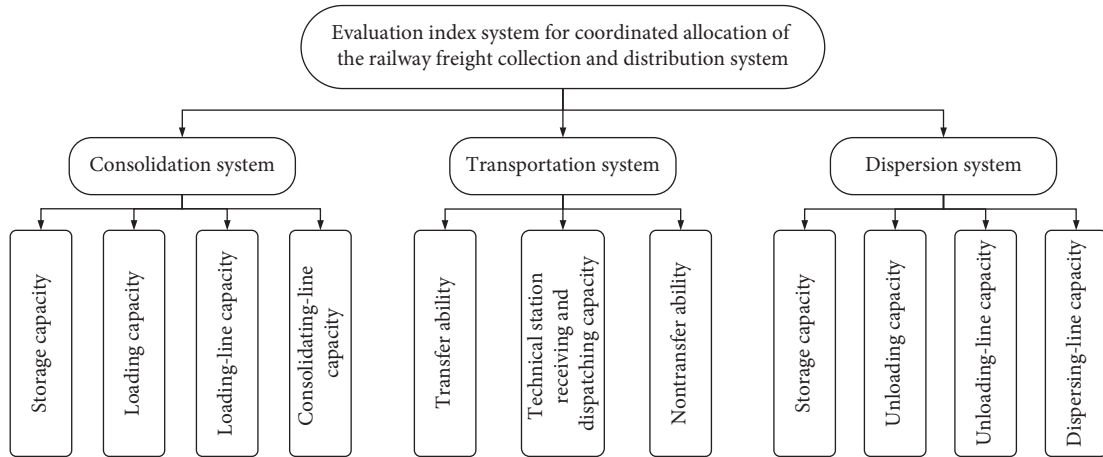


FIGURE 2: Evaluation index system for coordinated allocation of the railway freight collection and distribution system.

4. Entropy-TOPSIS Coupling-Coordinated Development Degree Model

This section introduces an “Entropy-TOPSIS coupling-coordinated development degree model” that the Entropy-weighted TOPSIS model and the coupling development degree model to quantify the synergy of the collection and distribution system. The proposed model can adapt to the real-time dynamic change of the railway network and realize the dynamic control of the collaborative configuration of the collection and distribution system. Simultaneously, for the ambiguity of the coordination size limit between the subsystems of the railway freight collection and distribution, we introduce the machine learning Fuzzy c-means (FCM) algorithm to clarify the intimacy and sparse relationship between the sub-samples, so as to divide the coordination size.

4.1. Entropy-TOPSIS Model. The Entropy-TOPSIS model [20] is an objective comprehensive evaluation method based on the finite unit multi-objective decision analysis in system engineering by combining the Entropy weight method with the TOPSIS model. The relative proximity, which is obtained

through calculation, reflects the overall situation of the coordinated development of the transportation system.

4.1.1. Entropy Weight Method to Determine Evaluation Index Weight

(1) *Data Standardization.* The data of each indicator are standardized. The indicators are generally divided into economic indicators and cost indicators.

For economic indicators,

$$y_{ij} = \frac{x_{ij} - x_{j\min}}{x_{j\max} - x_{j\min}} \tag{1}$$

For cost indicators,

$$y_{ij} = \frac{x_{j\max} - x_{ij}}{x_{j\max} - x_{j\min}} \tag{2}$$

where y_{ij} is the j index value of the i unit after dimensionless processing; x_{ij} is the original j index data of the i unit.

(2) Calculation of the Information Entropy of Each Evaluation Index

$$Y_{ij} = \frac{y_{ij}}{\sum_{i=1}^m y_{ij}},$$

$$Y = (Y_{ij})_{m \times n} = \begin{bmatrix} Y_{11} & \cdots & Y_{1n} \\ \vdots & \ddots & \vdots \\ Y_{m1} & \cdots & Y_{mn} \end{bmatrix}, \quad (3)$$

$$e_j = -\frac{1}{\ln m} \sum_{i=1}^m Y_{ij} \ln Y_{ij}, \quad j = 1, \dots, n.$$

where Y_{ij} is the j index value of the i unit after normalization processing, e_j is the information entropy value of the j index, e_j is not greater than 1, and $\ln m$ must be greater than 0.

(3) Calculation of the Weight of Each Evaluation Index

$$W_j = \frac{d_j}{\sum_{j=1}^n d_j}, \quad (4)$$

$$d_j = 1 - e_j. \quad (5)$$

where W_j is the weight of the j evaluation index; d_j is the information utility value. The smaller the entropy value e_j of the index, the larger the weight, which indicates that corresponding indicator carries more information in the integrated development of railway freight collection and distribution; otherwise, the less.

4.1.2. TOPSIS Evaluation Method to Determine Relative Proximity. The core of the TOPSIS evaluation method is to find the “relative proximity of the ideal point,” that is, to obtain the relative proximity between each evaluation object and the optimal solution, which is as a basis for evaluating the pros and cons.

(1) Establishing a Standardized Decision Matrix

$$r_{ij} = \frac{x_{ij}}{\sqrt{\sum_{j=1}^m x_{ij}^2}}, \quad 1 \leq i \leq m, 1 \leq j \leq n, \quad (6)$$

where x_{ij} is the j index value of the i unit; r_{ij} is the j index value of the i unit after the normalization process. Since the entropy weight method has obtained the normalized matrix, which is Y_{ij} , this step will not be repeated.

(2) Calculation of the Weighted Standardized Decision Matrix

$$V = [W_j Y_{ij}]_{m \times n}. \quad (7)$$

(3) Determining Positive and Negative Ideal Solutions

$$A^+ = \{v_1^+, \dots, v_n^+\},$$

$$A^- = \{v_1^-, \dots, v_n^-\}. \quad (8)$$

For economic indicators:

$$v_j^+ = \max\{v_{ij}, i = 1, \dots, m\},$$

$$v_j^- = \min\{v_{ij}, i = 1, \dots, m\}. \quad (9)$$

For cost indicators:

$$v_j^+ = \min\{v_{ij}, i = 1, \dots, m\},$$

$$v_j^- = \max\{v_{ij}, i = 1, \dots, m\}, \quad (10)$$

where A^+ is the positive ideal solution; A^- is the negative ideal solution.

(4) Calculation of the Distance from Each Unit to the Positive and Negative Ideal Solutions

$$D_i^+ = \sqrt{\sum_{j=1}^n (v_{ij} - v_j^+)^2}, \quad i = 1, \dots, m,$$

$$D_i^- = \sqrt{\sum_{j=1}^n (v_{ij} - v_j^-)^2}, \quad i = 1, \dots, m, \quad (11)$$

where D_i^+ is the distance from unit i to the positive ideal solution; D_i^- is the distance from unit i to the negative ideal solution.

(5) Calculation of the Relative Proximity of Each Unit to the Optimal Plan

$$O_i = \frac{\sqrt{\sum_{j=1}^n (v_{ij} - v_j^-)^2}}{\sqrt{\sum_{j=1}^n (v_{ij} - v_j^+)^2} + \sqrt{\sum_{j=1}^n (v_{ij} - v_j^-)^2}}, \quad i = 1, 2, \dots, m, \quad (12)$$

where O_i is the relative proximity. It takes a value between 0 and 1. The closer the O_i is to 1, the closer the evaluation object is to the optimal level; otherwise, the closer it is to 0, the closer the evaluation object is to the worst level. This paper uses relative proximity to express the comprehensive coordination of the system.

4.2. Coupling Development Degree Model

4.2.1. Coupling Degree. System coupling is used to indicate an interaction relationship between multiple systems (the number of systems ≥ 2). The coupling degree is used to characterize the degree of influence between subsystems. It is quantified by using the dispersion coefficient C_v in mathematics.

According to the definition of coordination degree, this paper constructs the comprehensive benefit function of the “collection and distribution” system :

$$\begin{aligned}
X &= f(x) = \sum_{j_1}^{n_1} W_j^1 Y_{ij}^1, \\
Y &= g(y) = \sum_{j_2}^{n_2} W_j^2 Y_{ij}^2, \\
Z &= h(z) = \sum_{j_3}^{n_3} W_j^3 Y_{ij}^3,
\end{aligned} \tag{13}$$

$f(x)$, $g(y)$, and $h(z)$ are the comprehensive benefit functions of the “consolidation,” “transportation,” and “dispersion” systems. The weight of each system index is obtained by formula (4).

As far as the binary system is concerned, taking the “consolidation system” and “transportation system” as examples, we hope that the smaller the deviation between $f(x)$ and $g(y)$, the better. That is,

$$C_v = \frac{S}{[f(x) + g(y)]/2} = \sqrt{2 \left\{ 1 - \frac{f(x) \cdot g(y)}{[(f(x) + g(y))/2]^2} \right\}}. \tag{14}$$

The smaller, the better (S is the standard deviation). As can be seen from the abovementioned formula, the necessary and sufficient conditions for making C_v smaller are better:

$$C = \frac{f(x) \cdot g(y)}{[(f(x) + g(y))/2]^2}. \tag{15}$$

The bigger, the better. We call the formula derived above as the “coupling coordinated degree model.” In the formula, C is called “coupling degree of collection and distribution subsystems” (also called coordination coefficient), which reflects the quantity and degree of combination and coordination between two subsystems.

4.2.2. Coupling Development Degree. Coupling degree, which is a quantitative index to measure the excellent degree of coordination between systems or elements, can be used to measure the degree of harmony and consistency among the systems or inner-elements during the development process. However, for different collection and distribution systems under the same index system, there may be situations in which the degree of coupling and coordination among subsystems or elements is extremely similar, but the actual coordination is not consistent. Hence, the coupling coordinated degree cannot fully and systematically describe the level of coordinated development between systems, leading to the loss of accuracy of the evaluation results. In this regard, in view of the connotation of coordinated development, based on the “Environment and Economic Coupling Development Degree Model” constructed by Liao [21] scholar, we construct a “Coupling Development Degree Model of railway freight collection and distribution” based on the content of this paper.

For the binary system, the “second-degree coupling” model (taking the “Consolidation-Transportation” system as an example) is built:

$$\begin{aligned}
D_{(2)} &= \sqrt{C_{(2)} \cdot T_{(2)}}, \\
C_{(2)} &= \frac{\sum_{j_1}^{n_1} W_j^1 Y_{ij}^1 * \sum_{j_2}^{n_2} W_j^2 Y_{ij}^2}{\left(\left(\sum_{j_1}^{n_1} W_j^1 Y_{ij}^1 + \sum_{j_2}^{n_2} W_j^2 Y_{ij}^2 \right) / 2 \right)^2}, \\
T_{(2)} &= \alpha * \sum_{j_1}^{n_1} W_j^1 Y_{ij}^1 + \beta * \sum_{j_2}^{n_2} W_j^2 Y_{ij}^2,
\end{aligned} \tag{16}$$

where $D_{(2)}$ is the coupling development degree, $C_{(2)}$ is the coupling degree of the binary systems, and $T_{(2)}$ is the overall benefit (or level) of the binary systems; α and β are undetermined coefficients. We believe that the importance of each link in the railway freight collection and distribution system is consistently, so the values of α and β are set to 0.5.

For the ternary system, the “three-degree coupling” model is built:

$$\begin{aligned}
D_{(3)} &= \sqrt{C_{(3)} \cdot T_{(3)}}, \\
C_{(3)} &= \left\{ \frac{\sum_{j_1}^{n_1} W_j^1 Y_{ij}^1 * \sum_{j_2}^{n_2} W_j^2 Y_{ij}^2 * \sum_{j_3}^{n_3} W_j^3 Y_{ij}^3}{\left[\sum_{j_1}^{n_1} W_j^1 Y_{ij}^1 + \sum_{j_2}^{n_2} W_j^2 Y_{ij}^2 + \sum_{j_3}^{n_3} W_j^3 Y_{ij}^3 \right]^3} \right\}^{1/3}, \\
T_{(3)} &= \alpha * \sum_{j_1}^{n_1} W_j^1 Y_{ij}^1 + \beta * \sum_{j_2}^{n_2} W_j^2 Y_{ij}^2 + \gamma * \sum_{j_3}^{n_3} W_j^3 Y_{ij}^3,
\end{aligned} \tag{17}$$

where $D_{(3)}$ is the coupling development degree, $C_{(3)}$ is the degree of coupling and coordination of the ternary system, $T_{(3)}$ is the overall benefit of the ternary system, α , β , and γ are the undetermined coefficients of the consolidation system, transportation system, and dispersion system, respectively, and $\alpha + \beta + \gamma = 1$. In this paper, the undetermined coefficients α , β , and γ are all attached with a value of 1/3.

The volume of the coupling coordinated development degree directly reflects the coordinated development degree of each subsystem or element in the system. The larger the value is, the stronger the close relationship of mutual cooperation and promotion exists between each subsystem or element, which is conducive to the sustainable development of the system.

4.3. Fuzzy Clustering FCM Algorithm

4.3.1. Algorithm Principle. The FCM algorithm is an unsupervised fuzzy clustering method based on the optimization of the objective function in the machine learning method. We use fuzzy mathematics to quantitatively determine the fuzzy relationship between samples quantitatively. So as to cluster objectively and accurately perform clustering, and divide the data set into multiple categories or clusters.

The mathematical relationship between the input and output of the FCM algorithm is shown as follows.

The input-Variable set is

$$X = \{x_1, x_2, \dots, x_n\}, \quad x_k \in R^p,$$

$$\text{output} \begin{cases} \text{Classification matrix: } U = \begin{pmatrix} u_{11} & \cdots & u_{1n} \\ \vdots & \ddots & \vdots \\ u_{c1} & \cdots & u_{cn} \end{pmatrix}_{c \times n}, \\ \text{Cluster center vector collection: } V = \{v_1, v_2, \dots, v_c\}, \quad v_k \in R^p. \end{cases} \quad (18)$$

where p represents the characteristics of the input elements, c is the number of clusters, and n is the number of elements in the data set. The cluster center represents the representative point of each class.

For the FCM algorithm:

Objective function:

$$J_m(U, V) = \sum_{i=1}^c \sum_{j=1}^n u_{ij}^m d_{ij}^2. \quad (19)$$

Restrictions:

$$\sum_{i=1}^c u_{ij} = 1, \quad 1 \leq j \leq n, \quad (20)$$

where U represents the original matrix, V represents the cluster center, u_{ij} is the degree of membership, which refers to the degree of membership of the j element corresponding to the i category; and d_{ij}^2 is the distance between the element j and the center point i under the Euclidean distance; m is a parameter of the degree of fuzzification. The interpretation of the restrictions is that the values of the degree of membership of an element to all categories add up to 1.

The mathematical relationship corresponding to the final effect to be achieved by clustering is to replace the objective function of FCM algorithm. The expression of the optimal solution is

$$\min(J_m(U, V)) = \min \left(\sum_{i=1}^c \sum_{j=1}^n u_{ij}^m d_{ij}^2 \right). \quad (21)$$

The Lagrangian multiplier method is used to construct function:

$$F = \sum_{i=1}^c \sum_{j=1}^n u_{ij}^m d_{ij}^2 + \sum_{j=1}^n \lambda_j \left(\sum_{i=1}^c u_{ij} - 1 \right). \quad (22)$$

By solving the objective function, the optimal solutions of U and V are obtained:

$$u_{ij} = \left[\sum_{k=1}^c \left(\frac{d_{ij}}{d_{kj}} \right)^{2/(m-1)} \right]^{-1},$$

$$v_i = \frac{\sum_{j=1}^n x_j u_{ij}^m}{\sum_{j=1}^n u_{ij}^m}. \quad (23)$$

In summary, the FCM algorithm requires two parameters, one is the number of clusters c and the other is the parameter m . In general, c is much smaller than the total number of cluster samples, and at the same time, $c > 1$ must be guaranteed. There is an evaluation function $L(c)$ for the selection of the number of clusters c :

$$L(c) = \frac{\sum_{i=1}^c \sum_{j=1}^n u_{ij}^m \|v_i - \bar{x}\|^2 / (c-1)}{\sum_{i=1}^c \sum_{j=1}^n u_{ij}^m \|x_j - v_i\|^2 / (n-c)}. \quad (24)$$

In the formula, the numerator represents the sum of the distances between classes, and the denominator represents the sum of the distances within the classes; m is a parameter that controls the flexibility of the algorithm. If m is too large, the clustering effect will be weak; if m is too small, the algorithm will approach the hard-clustering algorithm. By referring to the common sense of multiple papers, it is more appropriate to choose $m = 2.0$, which causes less noise to the data.

4.3.2. *Algorithm Steps.* According to the algorithm principle, the algorithm execution process is shown in Figure 3.

5. Case Analysis

Based on the research foundation of related projects, we select the Tongchuan Depot of Xi'an Railway Bureau that meets the basic requirements of the gathering and transportation system as the target system to start an example analysis.

The jurisdiction of the China Railway Xi'an Group Co., Ltd Tongchuan Depot is the two branch lines of Xiantong line and Meiqi line, with an operating mileage of 177.182 kilometers. The stations pass through Gaoling County, Sanyuan County, Yanliang District of Xi'an City, Fuping County, Yaozhou District of Tongchuan City, and Tongchuan City. It has 24 stations under its jurisdiction and is mainly responsible for the transportation of coal, aluminum, coke, cement, building materials, grain, chemical fertilizer, and other materials and people's daily necessities in the Tongchuan area, along the line and in various jurisdictions. The simplified plan of the Tongchuan Depot is shown in Figure 4.

Through the intelligent information system, the relevant guarantee information such as the capacity schedule of each loading and unloading point and the freight volume

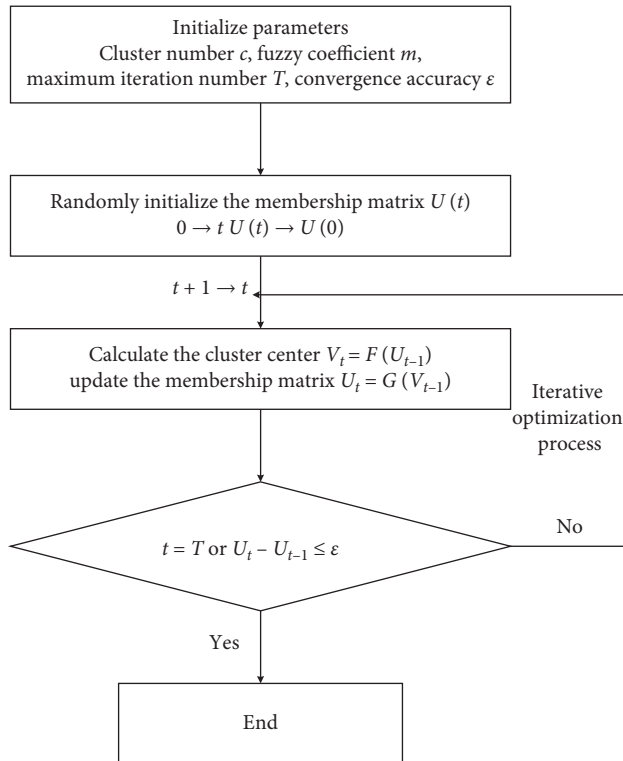


FIGURE 3: FCM algorithm flowchart.

summary report is extracted. By counting, analyzing, and filtering the information in the freight traffic report provided by the Tongchuan Depot, we select 13 stations as the target stations to build the collection and distribution system. After taking into account the expected difficulties in the subsequent indicators, we select the number of loading/unloading cargoes, the capacity of consolidating/distributing lines, the capacity of receiving/departing from technical stations, and the number of transfer or nontransfer cargoes as indicators. It should be noted that since the number of non-transfer cargoes at Meijiaping Station has decreased, the number of shunting cargoes can better represent the operating capacity of the technical station. Therefore, the analysis is mainly based on the number of shunting cargoes.

According to the selected index statistics, the initial data of the Tongchuan Train Depot collection and distribution system in 2017 is shown in Tables 1–3.

According to formula (12), the relative proximity of the system in each month of the Tongchuan Depot in 2017 is shown in Table 4.

It can be seen that in the first half of 2017, the coordination of the system fluctuates greatly, and the development is more irregular; in the second half, the coordinated transition is relatively small, and the development tends to be stable. Among them, the system in April has the highest comprehensive coordination and cooperates most closely coordinated system. This is closely related to the operation of each station every month.

In order to more intuitively reflect the closeness of the coordinated development among the various subsystems, based on the original data, according to the description in

Section 4, we construct a Multi-Entropy-TOPSIS Coupling Development Degree Mode. The obtained coupling development degree is a quantitative expression of the coordinated development relationship between the collection and distribution subsystems of the Tongchuan Depot in 2017 and can more intuitively reflect the strength of interaction and the level of coordinated development among the various subsystems. The obtained results are shown in Table 5. In order to more intuitively reflect the closeness of the coordinated development among the various subsystems, based on the original data, according to the description in Section 4, we construct a multi-Entropy-TOPSIS Coupling Development Degree Mode. The obtained coupling development degree is a quantitative expression of the coordinated development relationship between the collection and distribution subsystems of the Tongchuan Depot in 2017, and can more intuitively reflect the strength of interaction and the level of coordinated development among the various subsystems. The obtained results are shown in Table 5.

Through the horizontal analysis of results in Table 5, the coupling development degree between the binary systems is generally higher than that of the ternary system, indicating that the coordination between the binary systems is relatively tight, but the overall coordination is loose; in the results of the binary system coupling results D_2 , the “Consolidation-Dispersion” system coupling development degree is relatively low except for January and February, and the remaining months are at the highest values among other coupling methods. It shows that, among the various system coupling methods, the internal coordination between the consolidation system and dispersion system is the closest, which means that the operation efficiency of the “Consolidation-Dispersion” system is higher and the empty-weight distribution is more balanced.

Longitudinal analysis, as a whole, due to the increase in the number of passenger trains and the addition of temporary passenger trains during the Spring Festival, will relatively affect the development of freight capacity, so the coupling development degree between multiple systems is relatively low in February; from the range of the D_2 , in addition to “Consolidation-Dispersion,” the internal coordinated development of other multisystem combinations gradually changed from disorder to order. Among them, the value of the coupling coordination development degree in February and May is relatively low. We define them as the bottleneck months, and the impact factors involved are defined as bottleneck factors. Through the analysis of statistical data, it can be seen that the key factor affecting the month of bottlenecks is the Meijiaping Station, which is the hub technology station of the transportation subsystem. Its receiving-dispatching capacity and transfer capacity in February and May are among the lowest and second lowest in the year 2017.

Due to the difficulty of control, it is easier to control the binary system of the collection and distribution system under the same time and space than the ternary system. Hence, in view of the more difficult direct control, we consider improving the coordination of the ternary system by improving the coordination of the binary system to improve the overall coordination of the target transportation

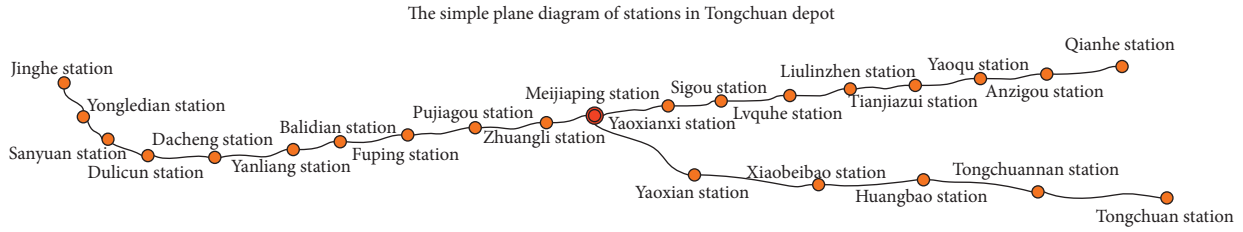


FIGURE 4: Simplified floor plan of the Tongchuan Depot.

TABLE 1: Initial statistics of the “consolidation” system of the Tongchuan Train Depot in each month in 2017.

Month	Qianhe loading capacity	Anzigou loading capacity	Yaoqu loading capacity	Tianjazui loading capacity	Liulinzhen loading capacity	Huangbao loading capacity	Yaoxian loading capacity	Consolidating-line1 capacity	Consolidating-line2 capacity
1	2979	1513	3103	1498	773	478	163	4121	634
2	3875	1497	4467	1663	1105	433	133	4689	493
3	3302	2113	3443	1942	1103	402	187	4002	513
4	4327	1892	4567	2384	1002	663	87	4896	576
5	4243	2189	4345	2415	1127	613	113	4890	553
6	3217	2113	3254	2267	833	601	98	3967	663
7	3174	1874	3665	2078	872	801	90	4323	830
8	3042	2298	4334	2575	1045	553	57	4897	590
9	3487	1684	3654	2288	788	153	32	4796	102
10	3198	1937	3471	2104	1015	133	34	4834	297
11	3023	1983	3112	1904	801	221	124	4089	404
12	3144	1873	3547	2075	843	234	81	4822	334

TABLE 2: Initial statistics of the “transportation” system of the Tongchuan Depot in 2017.

Month	Meijiaping receiving capacity	Meijiaping dispatching capacity	Meijiaping transfer ability
1	29167	29512	25780
2	21783	21914	18633
3	28694	28839	25510
4	26134	26048	23046
5	21976	20733	18696
6	24343	24121	20924
7	26196	25382	23142
8	23287	23490	20074
9	25334	25229	22013
10	27565	27475	24761
11	27432	27348	24001
12	28910	28905	25325

TABLE 3: Initial statistics of the “dispersion” system in the Tongchuan Depot in 2017.

Month	Fuping unloading capacity	Yanliang unloading capacity	Sanyuan unloading capacity	Yongledian unloading capacity	Jinghe unloading capacity	Dispersing-line capacity
1	11	111	366	27	552	674
2	12	39	221	18	421	784
3	23	24	401	25	572	853
4	12	58	532	32	601	835
5	5	62	640	47	653	894
6	16	94	620	102	731	889
7	33	68	756	98	551	857
8	20	95	715	0	482	898
9	56	150	988	31	445	1134
10	53	133	955	24	385	1032
11	85	86	717	12	477	986
12	95	63	555	27	455	779

TABLE 4: The relative proximity degree of the integrated collection and distribution system of Tongchuan Railway Depot in each month in 2017.

Month	O_i	Sort
1	0.3021	12
2	0.4470	4
3	0.4077	7
4	0.5674	1
5	0.5517	2
6	0.3900	10
7	0.4237	5
8	0.3733	11
9	0.4203	6
10	0.4045	8
11	0.4014	9
12	0.4592	3

TABLE 5: The results of the multiple coupling development degree D_i of the collection and distribution system of Tongchuan Depot in 2017.

Month	Binary system coupling D_2		Ternary system coupling D_3	
	Consolidation-transportation	Consolidation-dispersion	Transportation-dispersion	Consolidation-transportation-dispersion
	$D_{2,1}$	$D_{2,2}$	$D_{2,3}$	
1	0.1417	0.1349	0.1374	0.0797
2	0.0385	0.1192	0.0371	0.0492
3	0.1665	0.1610	0.1369	0.0921
4	0.1476	0.1785	0.1243	0.0951
5	0.0230	0.1927	0.0229	0.0519
6	0.1091	0.1887	0.1105	0.0837
7	0.1379	0.1957	0.1355	0.0937
8	0.0945	0.1662	0.0877	0.0753
9	0.1226	0.1829	0.1272	0.0870
10	0.1490	0.1815	0.1497	0.0937
11	0.1315	0.1588	0.1489	0.0864
12	0.1546	0.1807	0.1592	0.0960
Average \bar{D}_i	0.1180	0.1701	0.1148	0.0820

TABLE 6: Correlation coefficients between multiple systems.

Characteristic	Correlation coefficients
D_3 与 $D_{2,1}$	0.953
D_3 与 $D_{2,2}$	0.422
D_3 与 $D_{2,3}$	0.939

system. Using the Correl function, according to the data in Table 5, we calculate the correlation coefficient of the coupling coordination development degree of each multi-element system. The calculation results are shown in Table 6.

The results show that the coordinated development of the ternary system is not closely related to the ‘‘Consolidation-Dispersion’’ of the binary system, but highly related to the ‘‘Consolidation-Transportation’’ and ‘‘Transportation-Dispersion.’’ Therefore, if we want to improve the overall coordination of the constructed Tongchuan Depot collection and distribution system, the focus of the coordination configuration should be on strengthening the ‘‘Consolidation-Transportation’’ and ‘‘Transportation-Dispersion’’ of these two types of binary systems.

Combined with the above mentioned analysis, the bottleneck factors affecting the coordinated development of

TABLE 7: The relative proximity degree of the integrated collection and distribution system of Tongchuan Railway Depot in each month in 2018.

Month	O_i	Sort
1	0.5257	1
2	0.3615	12
3	0.4909	3
4	0.4252	7
5	0.4276	6
6	0.4645	4
7	0.5164	2
8	0.3766	11
9	0.3788	10
10	0.4473	5
11	0.3956	9
12	0.4133	8

Tongchuan Train Depot collection and distribution system and the correlation between multiple systems are analyzed. In order to meet the requirements of improving the volume of railway freight and the comprehensive utilization efficiency of the railway networks, this paper mainly focuses on

the following three aspects to strengthen the overall internal coordinated development of the collection and distribution system: the operation efficiency of the freight yard station, the traffic organization of the hub marshalling station and the technical standard of the station line to organize and control the coordinated configuration of the system in the 2017 bottleneck month of the Tongchuan Depot.

5.1. The Operation Efficiency of the Freight Yards and Stations. According to the results of the horizontal analysis of the coupling development degree, the system stations in January and February still need to strengthen the coordination ability of collecting and distributing operations. The operation efficiency of the freight collection and distribution terminal is mainly affected by the fluctuation of the transportation cargo flow, the fluctuation of the operation time, the unsmooth operation connection, the mutual interference and the interference of the adjacent subsystems. In view of the current freight environment of the Tongchuan Depot, it is recommended to make full use of the effective cargo space in the freight yard under its jurisdiction and strengthen the coordination of freight flow distribution. The number of cargos in the freight yard can not only directly affect the storage capacity of the yard, but also indirectly affect the loading and unloading capacity of the yard. According to “The table statistics of the capacity of each loading and unloading point in the Tongchuan Depot,” it can be considered that the Qianhezhen station, Anzigou station, Tianjiazui station and Liulinzhen station, all of which have less effective storage space, should adopt an effective storage space sharing strategy and strengthen the coordination of the distribution of empty and heavy cargoes flow among the freight yards. Secondly, to improve the collection and distribution efficiency, we consider strengthening the coordination efficiency of the internal operations of the coordinated collection and distribution yard, which can shorten the connection time.

5.2. Organization of Traffic Flow at a Hub Marshalling Station. According to the longitudinal analysis results of the coupling development degree, Meijiaping Station, which is the key marshalling station that affects the coordinated development of the entire “Consolidation-Transportation-Dispersion.” The poor organization of the operation coordination in the station directly causes the transit time of the cargoes and the residence time of the operation. The increase of negative feedback will affect the coordinated development of the whole railway system. For the inside of the yard, when the traffic flow arrives in a certain period of time or a direction, the yard should communicate with the dispatcher of the dispatching station to give priority access to the required traffic flow or nonadjustable transfer train. For urgent cargoes at the destination freight yard, the following principles should be adopted for coordination: “delivery first, picking up first; disassembling cares for delivery; delivery cares for

loading and unloading; loading and unloading cares for picking up; and picking up cares for marshling”.

5.3. The Technical Standards of Station Lines. At present, domestic freight stations are gradually fully electrified with single and double lines, but most of the freight yards in the Tongchuan Depot, such as Zhuangli Station, Yongledian Station and Yaoxian Station, still maintain manual loading and unloading, which will directly affect the operation efficiency and safety. In addition, as the modernization of railway logistics advances, it is necessary for the stations to upgrade the technical standards of the lines. It is possible to adopt reasonable configuration and transformation of the rain shed (quantity, construction and use area), platform, open cargo area, and special lines of some low-level technical freight yards in the Tongchuan Depot. By comparing the efficiency, capacity, price, technical parameters and other factors of facilities, we determine the facilities and equipment that should be configured in the freight yard, monitor the whole life cycle of facilities and equipment to realize dynamic facility equipment configuration and control the cost and improve the input-output ratio of the system.

Based on the abovementioned analysis, the bottleneck problems and rectification plans of the system are fed back to the Tongchuan Depot. After a year of follow-up statistics, at the end of 2018, the initial statistics of the collection and transportation system of the Tongchuan Depot under the same indicators were compiled. According to formula (12), the relative proximity of the integrated system of gathering and transportation in the Tongchuan Depot in 2018 is shown in Table 7.

Compared with the data in Table 4, it can be shown in Figure 5 that the overall coordination of the Tongchuan Depot collection and distribution system in 2018 has less fluctuations and a relatively smooth development compared with the 2017. It shows that after organizational adjustment, the overall development trend of coordination in each month of 2018 is relatively consistent and the overall development is relatively balanced.

In the same way, a multi-system Entropy-TOPSIS Coupling Development Degree Model is constructed to quantify the coordinated development of the collection and distribution subsystems of the Tongchuan Depot in 2018. The value of coupling coordination development degree is obtained as shown in Table 8.

The purpose of the paper is to analyze whether the synergy among the ternary subsystems of the Tongchuan Depot has been improved. However, as synergy is a fuzzy concept, the coupling development degree should be a quantified embodiment of the synergy. Therefore, we use FCM algorithm to divide the value of the coupling development degree to compare and analyze whether the coordination has been improved. For D_3 in 2017, first the number of clusters c by formula (24) is selected, and then it is determined that 4 is the optimal number of clusters for fuzzy clustering. The final clustering result is [(2, 5), (1, 8), (6, 9, 11), (3, 4, 7, 10, 12)]

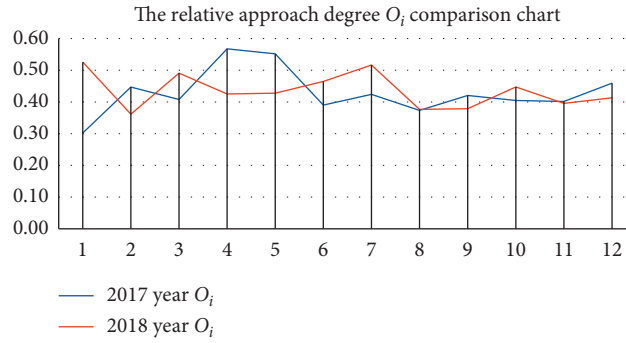


FIGURE 5: Comparison of relative proximity degree.

TABLE 8: The results of the multiple coupling coordinated development degree D_i of the collection and distribution system of Tongchuan Depot in 2018.

Month	Binary system coupling D_2		Ternary system coupling D_3	
	Consolidation-transportation $D_{2,1}$	Consolidation-dispersion $D_{2,2}$	Transportation-dispersion $D_{2,3}$	Consolidation-transportationdispersion
1	0.1721	0.1828	0.1563	0.0997
2	0.0405	0.1645	0.0401	0.0541
3	0.1717	0.1712	0.1436	0.0965
4	0.1407	0.1906	0.1376	0.0931
5	0.0499	0.1935	0.0495	0.0647
6	0.1100	0.1901	0.1139	0.0852
7	0.1527	0.2042	0.1500	0.1004
8	0.0990	0.1811	0.0970	0.0782
9	0.1262	0.1779	0.1274	0.0859
10	0.1600	0.1949	0.1507	0.0996
11	0.1469	0.1726	0.1455	0.0905
12	0.1598	0.1746	0.1549	0.0947
Average \bar{D}_i	0.1275	0.1832	0.1222	0.0869

TABLE 9: Levels of the coupling development degree of the ternary system in each month of 2017.

Month	Coupling development level
February and May	Primary coordinated development
January and August	Intermediate coordinated development
June, September, and November	Good coordinated development
March, April, July, October, and December	Quality coordinated development

(statistics by month). The classification results of the coupling development degree is shown in Table 9.

By using the level division in Table 9 as a reference, and by adopting the same calculation method in the above formula (24), we can determine that 3 is the optimal number of clusters for fuzzy clustering of D_3 in 2018, and the clustering result is [(2, 5), (4, 6, 8, 9, 11), (1, 3, 7, 10, 12)]. According to the classification of the coupling coordination level of the ternary system in each month of 2017, it can be seen that the value of the coupling development degree in each month of 2018 is higher than the highest value in the primary coordination development category, indicating that through the improvement of the bottleneck factors, the overall internal coordination of the system in 2018 is above the primary coordination; that is,

the coordination of the collection and distribution system has been improved.

In order to more intuitively reflect the coordinated floating changes between the multi-systems, the annual averages of the coordinated development degree of the binary and ternary systems are compared. The result is shown in Figure 6.

To sum up, through the analysis and improvement of the bottleneck factors that affect the coordinated development of the collection and distribution system, the following results can be found: for the overall development of the system, the system in 2018 is generally smooth and the synergy effect is better; for the internal coordination relationship of the system, from the comparison results of the coupling development degree, it can be seen that the coupling

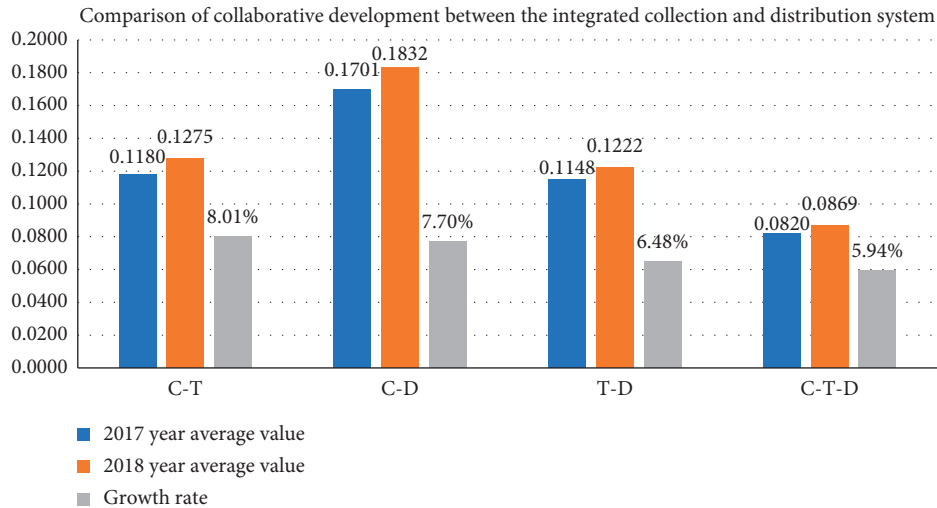


FIGURE 6: Comparison of the collaborative development between the integrated collection and distribution system.

coordination development degree between various multi-systems has improved compared with 2017. Furthermore, with the improvement of the capacity between the two types of binary systems of “Consolidation-Dispersion” and “Transportation-Dispersion,” the integration and coordination between the “Consolidation-Transportation-Dispersion” ternary subsystems of the Tongchuan Depot increased by 5.94%.

6. Conclusions

Based on the integration of collection and distribution theory, synergy theory, coupling theory, and fuzzy mathematics, this paper constructs the “Entropy-TOPSIS Coupling Development Degree Model” for quantitative analysis of the synergy between collection and distribution systems, which provides a new method for the research on the coordinated development of an integrated railway freight collection and distribution system. Compared with the expert scoring method and the AHP method, the Entropy-TOPSIS can help reduce the randomness and uncertainty of the overall collaborative evaluation of the system, and makes the evaluation method more reliable; the coupling development degree model can describe the internal coordinated development relationship of various aspects in detail, and the quantitative analysis effect is effective. The method simultaneously incorporates the ideas of fuzzy mathematics in machine learning, which has better robustness to the random freight flows. Through the quantitative analysis of the coordination of the Tongchuan Depot System of the China Railway Xi’an Group Co., Ltd. The bottleneck factors that affect the synergy development of the system are identified, and the rectification programs are proposed. Through one-year follow-up research, the results show that in 2018, the synergy of the Tongchuan Depot integrated transportation system in 2018 increased by 5.94% compared with that in 2017. The results verify the practicality of the intelligent analysis model. However, for

the weight of the model’s comprehensive benefit function, a part of the coupling development degree model in this paper, the current research assumes that the importance of each subsystem is the same, but in the actual transportation organization, the phenomenon of uneven resource allocation will inevitably occur. The importance of each system can be considered in subsequent studies, which will enrich the integrated research system of the railway freight collection and distribution. In addition, the analysis time will be refined, from month to week, so as to improve the effect of collaborative analysis.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This study was supported by the National Key R&D Program of China (2018YFB1201401) and the Fundamental Research Funds for the Central Universities (2019JBM030).

References

- [1] E. Alan, *Elements of Shipping*, Chapman & Hall, London, UK, 1989.
- [2] K. J. Button, *Transport Economics*, Edward Elgar Publishing Ltd., Cheltenham, UK, 2nd edition, 1990.
- [3] F. Huang, *Coordination and Optimization of Port Logistics System Collecting and Dispatching Links*, Southwest Jiaotong University, Chengdu, China, 2006.
- [4] Y. B. Geng, “Development of port collection and distribution railway under the strategy of transportation structure adjustment,” *Port Science & Technology*, vol. 12, pp. 2–5, 2019.

- [5] W. Xu, "Research on the construction scheme of special railway line for collection and distribution in Tonghai port," *Value Engineering*, vol. 14, pp. 176–179, 2020.
- [6] L. Y. Wu, *Research on the Coordination Evaluation of the Railway Transportation System in Caofeidian Port Area*, Beijing Jiaotong University, Beijing, China, 2018.
- [7] L. Kong and P. He, "Research on the capacity coordination of Shenhua heavy-duty coal transportation line gathering and transportation system," *Railway Transportation and Economy*, vol. 39, no. 13, pp. 57–62, 2018.
- [8] F. L. Fang, Z. Y. Chen, and Z. P. Lei, "Motivation and practice of synergetics for the integration of railway heavy-duty transportation collection and distribution," *Railway Transport and Economy*, vol. 8, pp. 25–29, 2018.
- [9] X. H. Yu, "Analysis of cooperative technology of gathering and distributing system of heavy haul railway," *Plant Maintenance Engineering*, vol. 10, pp. 152–153, 2019.
- [10] D. Liang and B. L. Lin, "Research on the strategic optimization model of railway transportation dynamic train flow organization," *Systems Engineering-Theory & Practice*, vol. 1, pp. 77–84, 2007.
- [11] L. Wang, B. L. Lin, J. J. Ma, and K. Y. Wen, "Methods for calculating dynamic traffic flow and optimizing transport resource allocation for bottleneck section of railway network," *China Railway Science*, vol. 3, pp. 116–123, 2016.
- [12] F. Xue, L. Zhao, and Q. L. Fan, "Research on coupling and adjustment of stage plan and dynamic car flow in marshalling station," *Journal of the China Railway Society*, vol. 1, no. 2, pp. 18–26, 2020.
- [13] R. Varadarajan, "Symbiotic marketing revisited," *Journal of Marketing*, vol. 50, no. 1, pp. 1473–1481, 1986.
- [14] H. Xie, "Analysis on the selection of collaborative marketing objects," *Modern Business Trade Industry*, vol. 22, no. 5, p. 121, 2020.
- [15] Y. L. Sun, *Research on Partner Evaluation in the Construction of Logistics Supply and Demand Alliance*, Northeastern University, Boston, MA, USA, 2006.
- [16] X. N. Yin and X. Z. Bao, "Financial risk assessment of emerging enterprises based on entropy weight TOPSIS method-taking biopharmaceutical industry as an example," *Friends of Accounting*, vol. 4, pp. 70–74, 2017.
- [17] Q. Wang and F. H. Tang, "Time and space differentiation of the coordinated development of ecological-economic-social system coupling in Dongting Lake area," *Economic Geography*, vol. 12, pp. 161–167+202, 2015.
- [18] Q. L. Zhang, G. Cheng, and J. Yang, "Research on the spatial connection of high-speed rail-population-economic coupling and coordination," *Science of Surveying and Mapping*, vol. 7, pp. 190–198, 2020.
- [19] H. K. Zhao, T. Wang, K. Song, and Y. S. Hu, "Research on the coordinated development of railway freight transportation system," *China Railways*, vol. 10, pp. 1–6, 2014.
- [20] W. C. Li, Z. X. Wang, and Q. X. Cui, "Design of multi-attribute group decision making based on five kinds of intuitionistic fuzzy numbers TOPSIS," *Statistics & Decision*, vol. 8, pp. 41–45, 2017.
- [21] C. B. Liao, "Quantitative Evaluation and classification system of coordinated development of environment and economy—taking the pearl river delta urban agglomeration as an example," *Tropical Geography*, vol. 19, no. 2, pp. 76–82, 1999.

Research Article

Research on the Simulation Application of Data Mining in Urban Spatial Structure

Jun Zhang, Xin Sui, and Xiong He 

School of Architecture and Urban Planning, Yunnan University, Kunming, Yunnan, China

Correspondence should be addressed to Xiong He; ydxh@mail.ynu.edu.cn

Received 17 May 2020; Revised 27 June 2020; Accepted 21 July 2020; Published 3 August 2020

Academic Editor: Naixue Xiong

Copyright © 2020 Jun Zhang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Data mining and simulation of the Internet of things (IOT) have been applied more and more widely in the rapidly developing urban research discipline. Urban spatial structure is an important field that needs to be explored in the sustainable urban development, while data mining is relatively rare in the research of urban spatial structure. In this study, 705,747 POI (Point of Interest) were used to conduct simulation analysis of western cities in China by mining the data of online maps. Through kernel density analysis and spatial correlation index, the distribution and aggregation characteristics of different types of POI data in urban space were analyzed and the spatial analysis and correlation characteristics among different functional centers of the city were obtained. The spatial structure of the city is characterized by “multicenters and multigroups”, and the distribution of multicenters is also shown in cities with different functional types. The development degree of different urban centers varies significantly, but most of them are still in their infancy. Data mining of Internet of things (IOT) has good adaptability in city simulation and will play an important role in urban research in the future.

1. Introduction

The concept of sustainable development was put forward in 1987, with the speeding up of globalization and urbanization; China's urban economy entered a period of rapid growth, the speed of urban development is accelerating and the scale of urban construction is generally increasing. However, at the same time, the urban problems brought about by the rapid development period have also attracted the attention of urban researchers. In the view of many scholars, the ideal sustainable urban spatial structure can alleviate or solve the problems that arise in the process of urban development in a limited way, which is an innovation of its mode and an unremitting exploration of the ideal urban spatial structure.

The development of the city presents a feature of continuous diffusion and reaggregation in space. In the process of aggregation and expansion in urban development, a gathering center will be formed in space. The gathering center is the space carrier of urban public activities and the core area of urban activities [1]. With the phenomenon of

diffusion and aggregation more obvious, the urban spatial structure has evolved from the original single center to a multicenter spatial structure [2, 3]. The multicenter spatial structure has quickly become a research hotspot and research difficulty in urban development research. Under the guidance of the current national macro policy and deliberate guidance of planners, megacities such as Beijing, Shanghai, Guangzhou, and Shenzhen have begun to use multicenters as the core goal of urban space development strategy [4–6].

Chinese scholars' research on the spatial structure of urban centers mainly involves urban center theory research [7, 8], central structure measurement and efficiency research [9–12], performance evaluation [8, 13–15], organization and governance [16, 17], etc. Most of China's domestic analysis is based on geospatial analysis and morphological analysis. The urban central structure is analyzed from the distribution characteristics of different elements of urban space [18, 19]. However, there are few studies on the spatial structure of urban centers from different functional types.

In recent years, the use of urban mass data to analyze the spatial structure of cities has provided a new research

paradigm for the sustainable development of cities. The use of a wider range of urban data includes rail transit card data, mobile phone signaling data, network review data, thermal map data, microblogging sign-in data, and night lighting data. Ying, Feng, Zhiqiang, Yang, and Xiong analyzed the spatial structure of the city through these “big data” cities [20–25]. However, based on the big data of these cities, there are problems in data acquisition difficulties, fewer ways, and inability to update in real time. Based on the open source data POI (Point of Interest), it provides an accurate and effective alternative to urban research. The POI data is an expression of a virtual abstraction of a real geographical entity in space, has spatial attribute information, has a large amount of data, and is easy to acquire. It is one of the most important data in the study of urban geography. It has been widely used in urban research. Chinese scholars’ research on POI data mainly focuses on urban spatial structure research [26], urban functional area identification [27], etc. Based on POI data, urban spatial structure can be well recognized, but through geographic information system and POI big data to cities. There is still little research on the development of the central space structure.

At present, the research on a large number of urban center spatial structures is mainly concentrated in first-tier cities and megacities such as Beijing, Shanghai, and Guangzhou. There is less attention to general provincial capital cities. Choosing China’s Guiyang as a research object is more representative. Guiyang is one of the core cities in western China and one of the birthplaces of China’s “big data.” In Guiyang’s urban sustainable development strategy, the development concept of building a multicenter city is also mentioned, and the spatial structure layout of the city’s multicenter is formed. To this end, this study uses the POI data of Guiyang City from 2016 to 2018, taking the main city of Guiyang as an example and using the nuclear density analysis and spatial correlation index in geography to analyze the evolution characteristics of urban centers in the main urban area of Guiyang. As one of the central cities in the west, Guiyang analyzes the spatial structure of the development of Guiyang multicenter. Through the analysis of Guiyang, it summarizes the achievements and shortcomings of the development of Guiyang urban center, and it provides new thinking for city planning and sustainable development mode of the city through the development of Guiyang urban center and the optimization of spatial structure.

Big data are used to simulate the urban spatial structure to achieve the purpose of exploring the urban spatial form of Guiyang, the capital city of western China. And further explore the sustainable development model of urban spatial structure.

2. Methodology

2.1. Study Area. Guiyang is the capital city of Guizhou province in western China and the central city in western China. By 2018, Guiyang had a permanent resident population of 3.136 million, with a GDP of US \$53.672 billion, ranking the second fastest growth rate in China. Since March 2015, cities in western China have been developing slowly

due to their inland location. The National Development and Reform Commission and other departments jointly issued the “One Belt And One Road” development strategy, bringing historical development opportunities to western cities. The urban spatial structure of Guiyang has undergone drastic changes (Figure 1).

2.2. Data Source. The data for this study was from December 2016 to December 2018. Get online map POI data about Guiyang, China, through an open interface. POI data has been widely used in urban navigation systems in urban geography, with spatial attributes and location information of urban entity objects. Therefore, the POI data basically contains all the entity objects in the study area urban space. As a kind of big data of geographic information, POI is widely used in the simulation of geographic space, especially in urban space, due to its advantages such as large quantity, fast update, and accurate positioning. Compared with other big data in model cities, POI has a better result.

After obtaining the POI data (<http://www.amap.com>), the data is verified and cleaned up. Among them, the 2016 POI data is 204,000, the 2017 POI data is 248516, and the 2018 POI data is 253231, a total of 705747. According to the online map POI classification system combined with different functional attributes of the city, 705,747 POI data are divided into five categories: public service, life service, leisure and entertainment, residence, and business (Table 1).

2.2.1. Kernel Density Estimation [28]. Kernel density analysis simulates the distribution of density by calculating the density values of points in the space and the neighborhood of the line features. In recent years, nuclear density analysis has been widely used in the study of spatial distribution of geography. This study compares the results of each nuclear density analysis under different search radii to explore the spatial distribution of the overall urban and urban areas of Guiyang’s main urban area and different types of POI:

$$P_i = \frac{1}{n\pi R^2} \times \sum_{j=1}^n k_j \left(1 - \frac{D_{ij}^2}{R^2}\right)^2, \quad (1)$$

where k_j is the spatial weight of the research object j (POI); D_{ij} is the distance between different POIs in space; and R is the bandwidth within the search area (bandwidth).

2.2.2. Spatial Correlation Index [29]. The spatial correlation index is currently used to describe the characteristics of cities in geospatial. This study uses the Getis-Ord General G and Getis-Ord G_i^* indices to measure the global and local features, structural patterns, and clustering of spatial locations in urban geospatial space, respectively. The spatial distribution of hot spots and cold spots is used to indicate the degree of association.

Getis-Ord General G :

$$G(d) = \frac{\sum \sum w_{ij}(d) x_i x_j}{\sum \sum x_i x_j}, \quad (2)$$

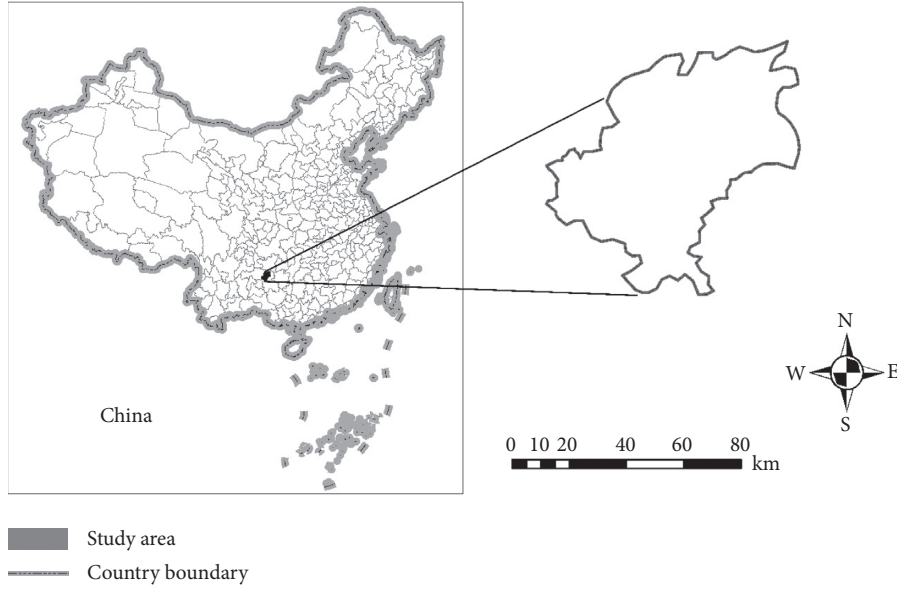


FIGURE 1: Study area.

TABLE 1: Classification of urban POI data in Guiyang.

POI classification	Classified content	Quantity (%)
Public service	Road auxiliary facilities, public facilities, transportation facilities services, access facilities, government agencies and social groups, and cultural and educational services	132117/18.72
Domestic services	Catering services, shopping services, and health-care services	435803/61.75
Leisure and entertainment	Resorts, golf-related, entertainment venues, and sports venues	49442/7
Reside	Community, villa area, and dormitory	26031/3.68
Business	Company, company, and bank	
Total		705747/100

where d is the distance of Guiyang city center; $w_{ij}(d)$ is the spatial weight in the study area; and x_i and x_j are the extended intensity index of urban land. Under the assumption that space does not agglomerate, the expected value of $G(d)$ is $E(G)$:

$$E(G) = \frac{W}{[n(n-1)]}, \quad (3)$$

$$w = \sum \sum w_{ij}(d).$$

Under the condition of positive distribution, the statistical test value of $G(d)$ is $Z(G)$:

$$Z(G) = \frac{[G - E(G)]}{\sqrt{\text{Var}(G)}}, \quad (4)$$

$$[E(G) = W_n(n-1)].$$

When $G(d)$ is higher than $E(G)$ and the $Z(G)$ value is significant, high-value clusters appear in the study area; when $G(d)$ approaches $E(G)$, the study area variables appear randomly distributed:

$$Z(G_i^*) = \frac{\sum_j^n w_{ij}(d)x_j}{\sum_j^n x_j}. \quad (5)$$

3. Result

3.1. Urban Center Evolution Process Based on Nuclear Density Analysis

3.1.1. Evolution of the Overall Urban Center. In nuclear density analysis, different search radii will result in different results for density analysis. The search radius of 500, 1000, 1500, and 2000 m was set to compare the density of POI in Guiyang main city from 2016 to 2018. It is found that, as the search radius increases, the internal POI will gradually merge. When the search radius is increased to a certain distance, the spatial agglomeration characteristics of the POI will be weakened to some extent. In general, when the search radius is small, a small range of POI aggregation areas can be identified. When the search radius is increased, it can reflect the macroscopic scale of the urban spatial structure, and the nuclear density equivalent curve will be smoother. This study explores the evolution of the urban center structure and

urban center. According to the scale comparison of existing urban research and the overall and local effects of different search distances in Kunming, 1500 m is finally selected as the nuclear density search radius of this study.

Based on the search radius $R=1500$ m, the nuclear density analysis of Guiyang three-year POI was carried out. It can be seen from Figure 2 that, in 2016, Guiyang City Center is Times Square, Longjiyuan and Tianjiao Haoyuan are subcenters, and Sunshine Garden and Ganjing Bridge have large groups. There are also obvious groups on both sides of Guanshan Lake. In 2017, Guiyang City Center was Times Square, but the city's subcenter changed, with Sunshine Garden surpassing Tianjiao Haoyuan and Longjiyuan as the city's subcenter. The group on both sides of the Ganjing Bridge and Guanshan Lake has an increasing trend, and a new urban group is formed in Longwan International. In 2018, the city center is still the Times Square. The subcenter of the city has undergone significant changes. The Longwan International Group has the fastest development, and the concentration is close to Longjiyuan, becoming a new subcenter of the city. The concentration of urban groups on both sides of the Ganjing Bridge and Guanshan Lake continues to increase (Figure 2).

Analysis of the POI nuclear density map of Guiyang City from 2016 to 2018, we can find that Guiyang has begun to have a multicenter spatial structure in the city since 2016. By 2018, the urban multicenter spatial structure has become more apparent. The main city of Guiyang is dominated by Times Square and has the widest range of radiation. It starts from Yunyan District in the north and goes to Guiyang Station in the south. In 2016–2018, there was a significant change in the city center of Guiyang. The main center of Guiyang City has been growing around Times Square. And it is very obvious to expand to Shantou Park. Guiyang City subcenter has a clear growth trend, but Tianjiao Garden is not growing faster than Sunshine Garden and Longwan International Group. There are many agglomeration groups in the main city of Guiyang, among which the development of Longwan International and Sunshine Garden Group is the most significant. The development of the groups on both sides of Guanshan Lake is slow, and the Ganshan Bridge also has obvious growth, but there is still a big gap from the city's subcenter.

Guiyang, China, is already developing into a multicenter city, but the development of the main and secondary centers is unbalanced. The growth of the main center in Guiyang is more obvious and the development is relatively perfect, but there is still a large space for the development of the subcenter and each group. In addition to the main center, the other secondary centers show the characteristics of dynamic growth. In general, the development of the spatial structure of the Guiyang urban center has initially highlighted the characteristics of dynamic “multicenter” development.

3.1.2. Evolution Process of Different Types of Urban Centers. Due to the urban development, planning policy guidance, and the public role of natural conditions, there are significant differences in the spatial distribution and agglomeration of urban centers with different functions. Therefore,

different types of POIs tend to have large differences in spatial distribution structures. The nuclear distribution of the centers of different functions is explored by performing nuclear density analysis on the distribution of different types of POIs.

As can be seen from Figure 3, the spatial distribution of different types of POIs is highly variable. In 2016, the public service category POI was centered on Times Square, and there were small groups in Guanji Lake Longjiyuan and Sunshine Garden. The residential POI is centered on Times Square, with Guiyang Station as the subcenter, and the main and subcenters have a tendency to merge with each other. There is a large group presence in Longwan International. The business category POI is centered on Times Square, and a small number of groups exist near Guanshan Lake. The life-oriented POI is centered on Times Square and spreads to the vicinity of Guiyang Station. Ganjing Bridge and Longjiyuan in Guanshanhu District are subcenters. Longwan International, Sunshine Garden, and Tianjiao Haoyuan have large urban groups. Leisure POI is centered on Times Square, and Longwan International is a group.

In 2017, the public service category POI was centered on Times Square, and the Longwan International Group had a growing trend. The residential POI is centered on Times Square and the subcenter has been integrated. And they grew to Guiyang Station and Shenlong Cave, respectively. The Guanji Lake Longjiyuan Group continued to grow. The business category POI is centered on Times Square and Shantou Park is the subcenter. Longjiyuan has a group presence, which is a big change from 2016. The life class POI is centered on Times Square. The radiation area has further increased, and the concentration of Longwan International, Sunshine Garden, and Tianjiao Haoyuan City has caught up with the Ganjing Bridge and Longjiyuan in Guanshanhu District. The same as the city's subcenter, a new city group appeared on both sides of Guanshan Lake. The concentration of leisure POI urban main centers and urban groups has further increased.

In 2018, the public service POI was centered on Times Square, and the growth trend of the main center was more obvious than the growth trend of the group. The residential POI is centered on Times Square, and the growth trend toward Guiyang Station and Shenlong Cave is more obvious. The business category POI is centered on Times Square, and the Shantou Park subcenter tends to merge with the main center. In the vicinity of Guanshanhu Park, the phenomenon of urban group formation is more obvious. The life class POI is centered on Times Square. A new subcenter was formed near Guiyang Station, and the concentration of POI in the other subcenters was also more obvious. The leisure POI city main center has a differentiation trend in the direction of Shenlong Cave.

From the analysis of different types of POI nuclear density maps from 2016 to 2018, we can find that, in the past three years, different types of POIs are based on Times Square. The development of urban centers with different functions is quite different. Among them, the development of living services and residential urban centers is relatively perfect, but the center of life services is growing faster than leisure centers.

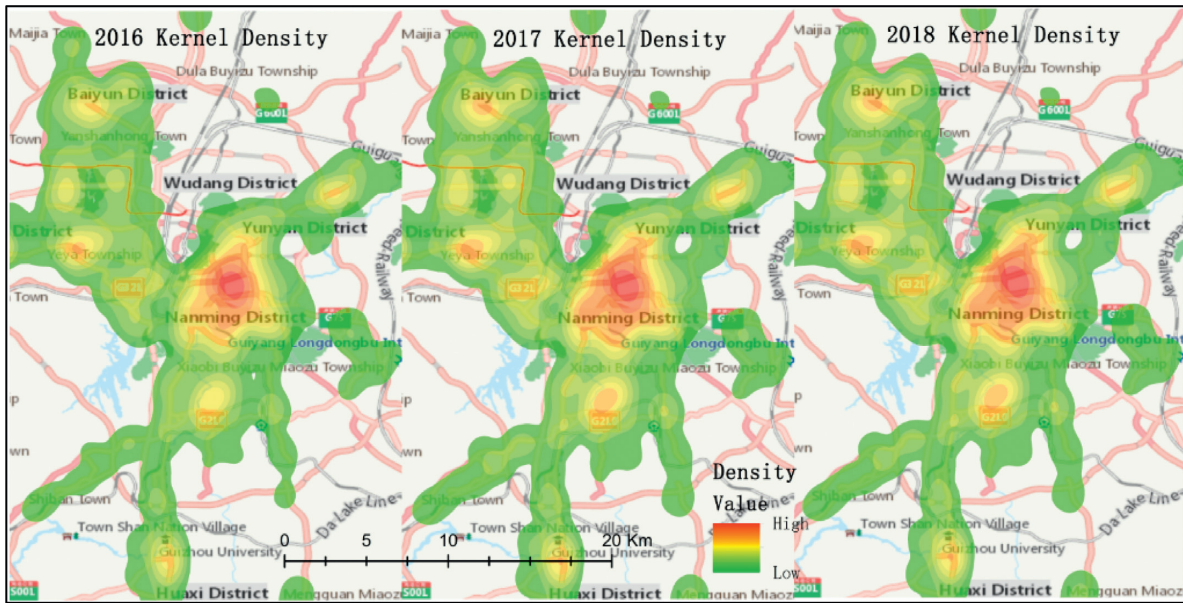


FIGURE 2: POI nuclear density from 2016 to 2018.

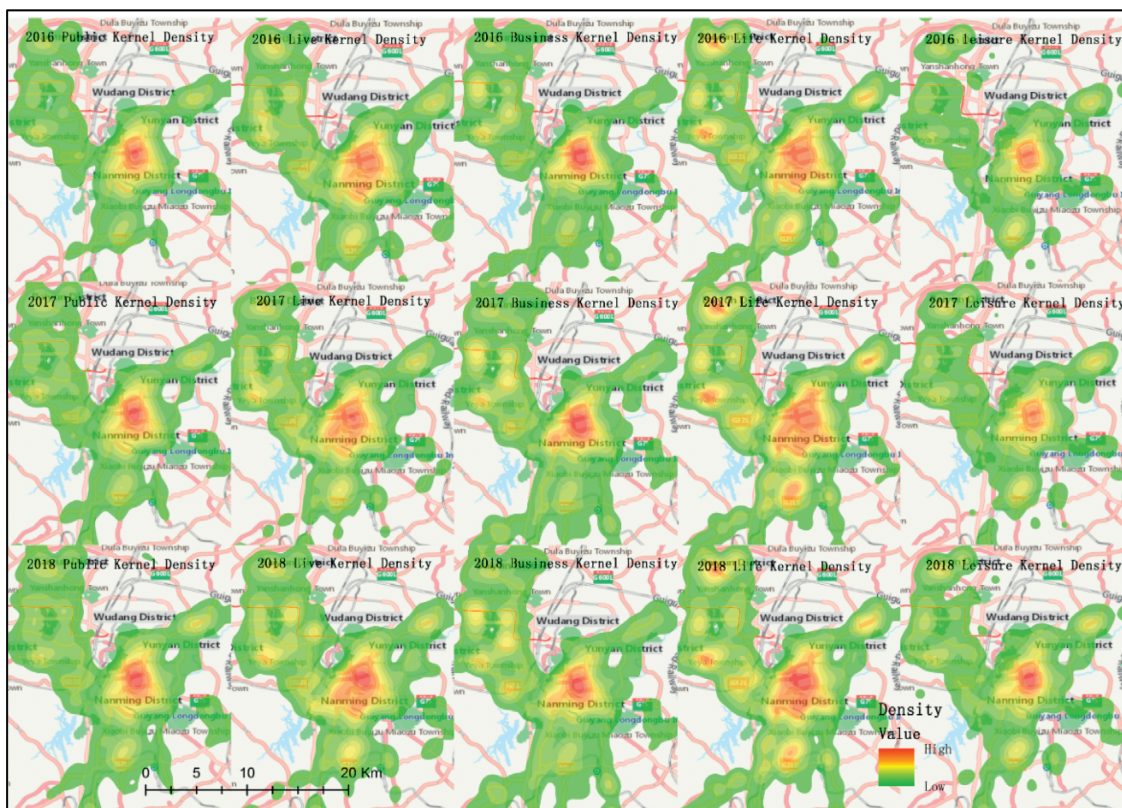


FIGURE 3: Nuclear density of different types of POIs from 2016 to 2018.

The public service and business city centers are generally developed. The growth rate of the main center is much larger than that of the subcenters and groups. Leisure basically exists in the form of a single center.

There are great imbalances in the development of different functional city centers in Guiyang, China. Some urban centers only have one single center, while the degree of

development of the main centers in other cities is much higher than that of the secondary centers and clusters. There is even a situation that the secondary centers and clusters are assimilated into the main centers with the development of the main centers. The multicenter of urban function is still in its infancy, and the spatial structure of urban center also presents an obvious dynamic change.

3.2. Urban Center Evolution Based on Spatial Correlation Index. Based on the evaluation unit of Guiyang City and the degree of urban POI aggregation, the spatial growth correlation index of the city center of Kunming, Getis-Ord General G and Getis-Ord G_i^* , is calculated. Thus, the global and local characteristics of the city center due to spatial correlation are described.

The global characteristics are shown in Table 2: in three years, the observation value $G(d)$ of the global statistical indicators of the Guiyang City Center Growth Correlation Index is greater than the expected value $E(d)$, and the Z score is obvious, indicating that it is in the main city of Guiyang. There is a phenomenon of high/low value agglomeration, but the Z value rises and falls, and the agglomeration phenomenon also increases and decreases. The overall Z value is the highest in 2016–2018, indicating that the growth of POI in Guiyang has been growing around a few cores in the past three years. Except for the overall Z value, the Z values of other functional POIs have increased, and the central polarization has been strengthened; 2016–2018 The public service, living service, and business Z value are relatively high, indicating that the growth of three types of POIs around several centers in three years is obvious; in 2016–2018, the Z value of public services and life services has increased rapidly, and the central polarization has strengthened. And the Z value of life service rose the fastest, indicating that the polarization intensity of the high/low agglomeration center is rapidly increasing. The above analysis also verified the results of the nuclear density analysis.

The spatial correlation index Getis-Ord G_i^* in the three-year evaluation unit is calculated separately, and the high-level cartographic representation based on ArcGIS is used to obtain the “hot spot area” and “cold spot area” distribution map of the POI growth intensity in the main urban area of Guiyang (Figure 4). The hot spot area and the “cold spot area” have obvious evolution and migration processes in the spatial distribution.

Analysis of the distribution map shows that, from 2016 to 2018, with the acceleration of the urbanization process in Guiyang, the urban POI increased, the number of “hot spots” increased significantly, and the “cold spot area” decreased continuously. It has risen from 33.6% in 2016 to 38.2% in 2017 and reached a maximum of 40.8% in 2018. In the three periods, the “cold spot zone” became a significant downward trend, accounting for 33.9%, 29.7%, and 24.8%, respectively. In the three periods, except for Times Square, which has been in the “hot spot”, “hot spot” such as Longwan International, Longjiyuan, and Sunshine Garden has been constantly evolving and changing.

In 2016–2018, there was no obvious change in the “hot spot” of public service category, mainly due to the increase in the number of hotspots and cold spots, and the growth of the hotspots as Times Square, but there were also new hotspots in Longwan International.

In 2016–2018, the “hot spot” of the residential category changed more obviously. First of all, the “hot spot” increased to 69.7% within three years. Secondly, outside the Times Square, new growth cores appeared in Longwan International, Longjiyuan, and Sunshine Garden. The growth

nuclear periphery was followed by high-value area and median area and low-value area; the peripheral structure is more obvious.

In 2016–2018, the change of “hot spot” in business category mainly showed that it changed from south of Times Square to the north, and more “cold spot areas” were converted into median areas, but the increase of “hot spots” was not obvious.

The “hot spot” of life service category has the most significant changes. First, the “cold spot area” is all converted into “hot spot area” and “media area”, and then the “hot spot area” is spread to the whole city, in Times Square, Longwan International, Guanjiu District Longjiyuan, and Sunshine Garden have all become the core of growth and have reached the development level of urban multicenter.

The leisure “hot spot” has not changed significantly in three years, and its growth rate is slow.

According to the above analysis, in 2016–2018, the growth characteristics of various POIs in Guiyang City showed a rapid growth process, and the “hot spot” was roughly reflected in the development process of the upper and lower ends of Times Square. In 2016–2018, various POIs are mainly concentrated in Times Square, which guides the growth of Guiyang City Center. During the continuous expansion of the “hot spot,” it gradually evolved into a relatively continuous gathering area in the Guanshan Lake area, and it tends to develop in the direction of Longjiyuan in Guanshanhu District. In the three-year change, basically the spatial structure of the single center of Times Square has been transformed into a spatial structure led by Times Square, Longwan International, Guanshan Lake, Longjiyuan, and many other city centers.

It can be seen from the spatial correlation index Getis-Ord General G and Getis-Ord G_i^* that the spatial structure change of the urban center of Guiyang is a very obvious dynamic “multicenter” development process. However, except the main urban centers, the other centers were not significantly developed. With the passage of time, the sustainable development model of the urban center spatial structure in Guiyang is bound to be a dynamic “multicenter” development.

4. Discussion

4.1. Simulation of Urban Center Evolution Based on POI Data.

This study studied the spatial structure of urban centers based on the analysis of POI data from 2016–2018 through geographic information technology. For the perspective of urban geography research and the research paradigm, the POI can describe in detail the geographical distribution and agglomeration of various elements in the city and then identify the geospatial structure of the city. Compared with traditional conventional data, POI data has better observability and timeliness. The clustering characteristics of POI can quickly reflect the problems of urban geospatial construction, industrial distribution, and functional improvement. Therefore, it provides a strong reference for urban development. However, the POI data is only the abstract expression of the urban entity in geospatial space, lacking the

TABLE 2: Getis-Ord General G index of POI in Guiyang main city from 2016 to 2018.

		Total	Public service	Domestic services	Leisure	Live	Business
2016	$G(d)$	0.000087	0.000174	0.000159	0.000180	0.000203	0.000246
	$E(d)$	0.000019	0.000077	0.000096	0.000117	0.000155	0.000110
	Z	154.765599	49.719755	31.666177	18.253169	18.999336	35.250216
2017	$G(d)$	0.000078	0.000169	0.000137	0.000165	0.000220	0.000198
	$E(d)$	0.000018	0.000079	0.000080	0.000114	0.000174	0.000089
	Z	154.646603	54.024563	36.362221	17.605138	18.541761	41.195000
2018	$G(d)$	0.000072	0.000133	0.000004	0.000165	0.000214	0.000174
	$E(d)$	0.000018	0.000055	0.000002	0.000104	0.000165	0.000085
	Z	156.753601	70.542463	146.959363	22.548152	20.214073	41.786830

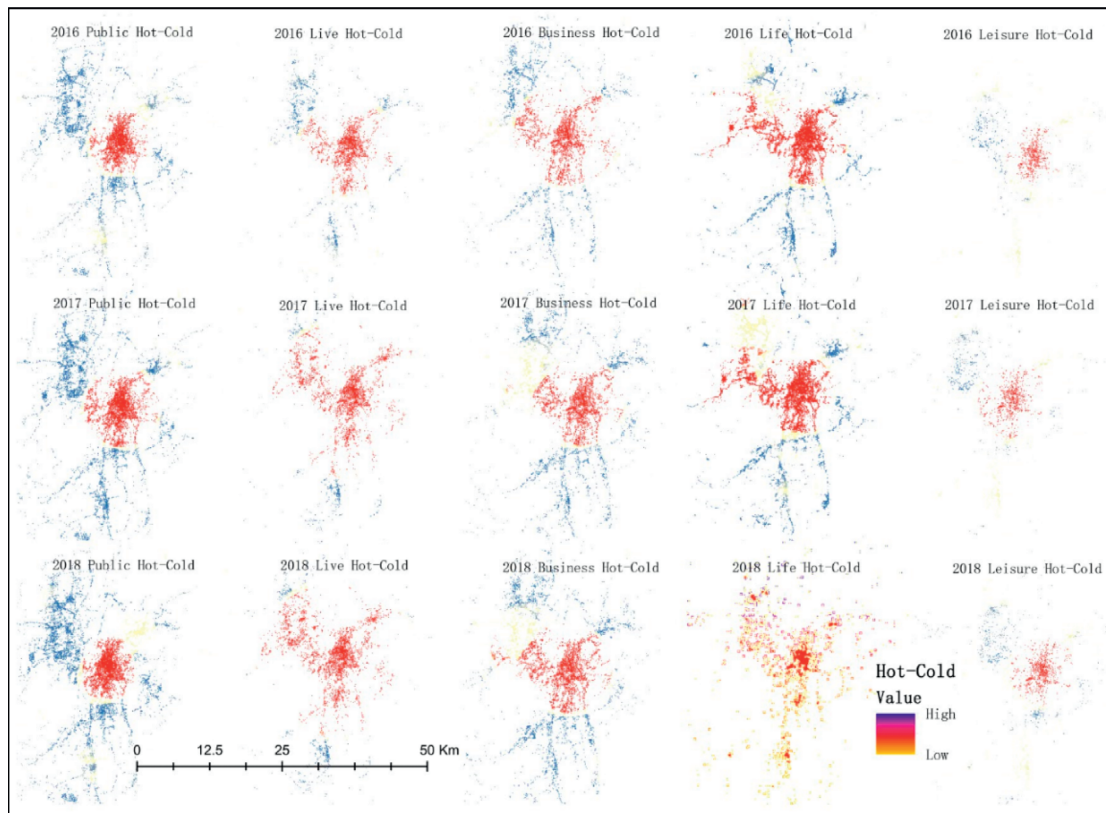


FIGURE 4: The evolution of the “hot spot” in the main urban area of Guiyang from 2016 to 2018.

detailed information of the attributes of the geographic entity. This study does not consider the grade and scope of the POI in the urban space. In the future, the spatial weight of the POI will be deeply studied.

4.2. The Wide Application of Big Data to Simulate Urban Space. It can be seen from this study that big data has been playing an increasingly important role in urban spatial computing. In today's development of new disciplines, traditional data in urban simulation computing can no longer adapt to the development of new disciplines. More and more big data should be introduced into urban simulation calculation, among which POI data is a typical one. POI big data has been used to simulate the calculation of cities with good results. In future studies, more and more

updated big data should be introduced to calculate cities in order to obtain better results.

4.3. Prospects for the Integration of Geographic Information Technology and Urban Geographic Big Data. The spatial structure of urban centers based on sustainable development belongs to the discussion of the sustainability of urban space. The research on the sustainability of urban space is a hot issue in urban development, and it is also an urgent problem needed to be solved. The sustainable development model of urban spatial structure is closely related to urban nature, urban population, land use scale, infrastructure, and other factors. This study studied the spatial structure distribution of urban centers through the analysis of urban big data and analyzed that the sustainable development mode of Guiyang

urban space is a dynamic “multicenter” development, which provides a good reference for the sustainable development of urban space structure in terms of exploration methods and results.

5. Conclusion

This study is based on POI data, using nuclear density analysis and spatial correlation index, according to the distribution, agglomeration, and correlation characteristics of POI data in space. The evolution characteristics of different urban centers in the main urban area of Guiyang City were discussed, and the following preliminary conclusions were obtained.

Guiyang’s main urban area presents a more obvious “multicenter, multigroup” space structure, and Times Square dominates Guiyang’s urban development.

Under five different types of functions, the urban structure has a relatively obvious multicenter distribution. Public services, living services, residence, and business functions are relatively complete. The public service and life service center are mature. In particular, the Life Service Center is basically distributed in multiple centers.

In general, the development of the central areas of Guiyang’s main city is relatively perfect. Although the multicenter development of the city is still in its infancy, the differences between the centers are obvious, and there is a big gap from the multicenters of the city. However, the centers in Guiyang are independent of each other and have great development prospects. In the future, the spatial structure of urban center in Guiyang will inevitably present a sustainable development model of dynamic “multicenter” development.

The urban spatial structure and form explored through big data is more authentic and objective in the expression of cities. It plays a guiding and leading role in the future urban planning and construction and provides references for urban builders and planners.

Data Availability

All raw data in this study are available free of charge. Readers who wish to repeat this study can do so through the following link (DOI 10.17605/OSF.IO/MP29E).

Conflicts of Interest

The authors declare no conflicts of interest.

Authors’ Contributions

Z. J. conducted a theoretical analysis and contributed to drafting the first draft and writing. X. H. improved the experiment and argumentation and scientifically demonstrated the results of the experiment. X. S. has carried out a lot of work improvement on the writing of the article and participated in the experiment.

References

- [1] J. Yang, Z. Biao, and Y. Shi, “On theoretical frameworks of urban center system development,” *Urban Planning Forum*, vol. 1, pp. 33–39, 2012.
- [2] S. W. Sun, “City center and city public space: a planning review on the construction of Lujiazui area of Pudong district, Shanghai,” *City Planning Review*, vol. 8, pp. 66–74, 2006.
- [3] X. W. Zheng, “Identification and optimization of xi’an urban center system based on open data,” *Planners*, vol. 31, pp. 57–64, 2017.
- [4] L. N. Gao, “Study on the relationship between polycentric development and economic performance of urban agglomeration—a case study of Yangtze river delta mega-city region,” *Science & Technology Progress and Policy*, vol. 35, pp. 46–52, 2018.
- [5] J. X. Nie, Y. P. Huang, and Z. R. Shan, “Characteristics and formation mechanism of the urban system of Wuhan metropolitan area: a study based on the perspective of urban network,” *Modern Urban Research*, vol. 12, pp. 110–116, 2018.
- [6] F. Q. Li, M. Zhao, M. D. Wu, and J. Z. Huang, “Polycentric mega-city and its mechanism of spatial performance: findings from xiamen based on LBS and census data,” *Urban Planning Forum*, vol. 5, pp. 21–32, 2017.
- [7] Z. D. Luo and C. S. Zhu, “Understanding polycentricity by configuration, function and governance,” *Urban Planning International*, vol. 23, pp. 85–88, 2008.
- [8] Y. P. Wei and M. Zhao, “Spatial structure and performance of metropolis: interpretation and application of polycentric structure,” *City Planning Review*, vol. 4, pp. 9–16, 2006.
- [9] B. D. Sun, T. Tu, W. Shi, and Y. L. Gao, “Test on the performance of polycentric spatial structure as a measure of congestion reduction in megacities: the case study of Shanghai,” *Urban Planning Forum*, vol. 2, pp. 63–69, 2013.
- [10] C. R. Ding, “The impact of urban spatial structure and land use pattern on urban transportation,” *Urban Transport of China*, vol. 8, pp. 28–35, 2010.
- [11] Y. L. Guo, *Study on the Influence of Multi-Center Space Structure on Urban Traffic*. Master’s Degree, East China Normal University, Shanghai, China, 2011.
- [12] K. Yang, “Population distribution and multicenter measurement of great Beijing,” *China Population Resources and Environment*, vol. 25, pp. 83–89, 2015.
- [13] F. Q. Li and M. Zhao, “A discussion on housing development in multi-center cities: phenomenon & planning implications of Chongqing,” *Urban Planning Forum*, vol. 3, pp. 8–19, 2011.
- [14] X. H. Wang and B. D. Sun, “The economic performance of the polycentric spatial structure of mega-cities: based on the models of urban economics,” *Urban Planning Forum*, vol. 22, pp. 20–27, 2011.
- [15] H. Yan and B. D. Sun, “The impact of polycentric urban spatial structure on energy consumption: empirical study on the prefecture-level and above cities in China,” *Urban Development Studies*, vol. 22, pp. 13–19, 2015.
- [16] A. G. Liu and K. Z. Yang, “Comments on Krugman’s edge city model,” *Scientia Geographica Sinica*, vol. 21, pp. 315–322, 2001.
- [17] J. X. Zhang, X. L. Luo, and J. Yin, “Polycentric mega-city regions and multi-level governance of the yangtze river delta,” *Urban Planning International*, vol. 23, pp. 65–69, 2008.
- [18] B. D. Sun and X. H. Wei, “Spatial distribution and structure evolution of employment and population in Shanghai Metropolitan Area,” *Acta Geographica Sinica*, vol. 69, pp. 747–758, 2014.

- [19] L. Zhang, W. Z. Yue, and Y. Liu, "Multidimensional analysis of the polycentric urban spatial structure—a case of Hangzhou," *Economic Geography*, vol. 37, pp. 67–75, 2017.
- [20] Y. Long, Y. Zhang, and C. H. Cui, "Identifying commuting pattern of Beijing using bus smart card data," *Acta Geographica Sinica*, vol. 67, pp. 1339–1352, 2012.
- [21] F. Zhen, B. Wang, and Y. X. Chen, "China's city network characteristics based on social network space: an empirical analysis of sina micro-blog," *Acta Geographica Sinica*, vol. 67, pp. 1031–1043, 2012.
- [22] Z. Q. Wu and Z. N. Ye, "Research on urban spatial structure based on baidu heat map: a case study on the central city of Shanghai," *City Planning Review*, vol. 40, pp. 33–40, 2016.
- [23] J. Cai, B. Huang, and Y. Song, "Using multi-source geospatial big data to identify the structure of polycentric cities," *Remote Sensing of Environment*, vol. 202, pp. 210–221, 2017.
- [24] X. Yang, D. Wang, and F. Jia, "Exploring the disparities in park access through mobile phone data: evidence from Shanghai, China," *Landscape and Urban Planning*, vol. 181, pp. 80–91, 2019.
- [25] X. He, Z. Yang, K. Zhang, P. Yang, and S. Zhang, "The spatial distribution patterns of the catering trade in nanchang based on Internet public reviews," *International Journal of Technology*, vol. 9, no. 7, pp. 1319–1328, 2018.
- [26] B. C. Ding, Y. X. Liu, and G. Chen, "Urban spatial structure of port city in South China sea based on spatial coupling between nighttime light data and POI," *Journal of Geo-Information Science*, vol. 20, pp. 854–861, 2018.
- [27] Z. D. Chen, B. W. Qiao, and J. Zhang, "Identification and spatial interaction of urban functional regions in beijing based on the characteristics of residents' traveling," *Journal of Geo-Information Science*, vol. 20, pp. 291–301, 2018.
- [28] C. Otioma, A. M. Madureira, and J. Martinez, "Spatial analysis of urban digital divide in Kigali, Rwanda," *GeoJournal*, vol. 84, no. 3, pp. 719–741, 2019.
- [29] Y. Wang, X. Li, Y. Kang, W. Chen, M. Zhao, and W. Li, "Analyzing the impact of urbanization quality on CO2 emissions: what can geographically weighted regression tell us?" *Renewable and Sustainable Energy Reviews*, vol. 104, pp. 127–136, 2019.