

Volume 7, Issue 2, 2025

Print ISSN: 2663-1024

Online ISSN: 2663-1016

EURASIA JOURNAL OF SCIENCE AND TECHNOLOGY



Copyright© Upubscience Publisher

Eurasia Journal of Science and Technology

Volume 7, Issue 2, 2025



Published by Upubscience Publisher

Copyright© The Authors

Upubscience Publisher adheres to the principles of Creative Commons, meaning that we do not claim copyright of the work we publish. We only ask people using one of our publications to respect the integrity of the work and to refer to the original location, title and author(s).

Copyright on any article is retained by the author(s) under the Creative Commons

Attribution license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Authors grant us a license to publish the article and identify us as the original publisher.

Authors also grant any third party the right to use, distribute and reproduce the article in any medium, provided the original work is properly cited.

Eurasia Journal of Science and Technology

Print ISSN: 2663-1024 Online ISSN: 2663-1016

Email: info@upubscience.com

Website: <http://www.upubscience.com/>

Table of Content

THEORY AND APPLICATIONS OF CRIME SITUATION AWARENESS TECHNOLOGY LiNing Yuan, ZhongYu Xing*, YuXia Tang	1-6
A-SHARE INTELLIGENT STOCK SELECTION STRATEGY BASED ON THE DEEPSEEK LARGE MODEL: TECHNICAL ROUTES, FACTOR SYSTEMS, AND EMPIRICAL RESEARCH HaiLong Liao	7-13
CUSTOMER SEGMENTATION AND CHURN PREDICTION BASED ON K-MEANS AND RANDOM FOREST: A CASE STUDY OF E-COMMERCE DATA ZhuoRan Li	14-19
THE OPTIMAL YIELD PROBLEM OF CROP PLANTING BASED ON LINEAR PROGRAMMING MODEL Jie Deng, Jian Wang, Hao Xu, ZhengBo Li*	20-27
THE CROP PLANTING PROBLEM BASED ON THE OPTIMIZATION MODEL YiTong Liu*, YiLin Wang, JiLing Zou	28-33
MATCHING CONTROL STRATEGY FOR HYDROGEN CIRCULATION SYSTEM FOR ON-BOARD FUEL CELLS JianHua Liu*, Jun Li, JingGuang Xie, NanNan Liao, YaNan Gao	34-46
– 30 ℃ COLD START STRATEGY OF DESIGNED FUEL CELL SYSTEM YinHao Yang*, JueXiao Chen, Chang Du	47-54
CONSTRUCTION OF CLASSIFICATION METHOD FOR URBAN ROAD INTERSECTIONS Lai Wei, XiChan Zhu, ZhiXiong Ma*	55-62
STREETSCAPE GREENNESS AND PARK SERVICE EVALUATION IN ZHENGZHOU, CHINA: A SPATIAL MULTI-ZONING PERSPECTIVE Da Mao*, ZhiYu Yuan, MengLei Zhao, HuaYu Fu	63-71
PERFORMANCE COMPARISON OF FREE-PISTON STIRLING CRYOCOOLERS WITH METALLIC AND NON-METALLIC PACKING IN WOUND REGENERATORS ShuLing Guo, AnKuo Zhang*	72-83

THEORY AND APPLICATIONS OF CRIME SITUATION AWARENESS TECHNOLOGY

LiNing Yuan, ZhongYu Xing*, YuXia Tang

School of Information Technology, Guangxi Police College, Nanning 53028, Guangxi, China.

Corresponding Author: ZhongYu Xing, Email: xingzhongyu@gxjxcxy.edu.cn

Abstract: Crime situation awareness technology utilizes crime big data to analyze the patterns and trends of criminal incidents, predict potential future crimes, and promote the transformation of public safety towards proactive prevention. It provides decision-making support for public security such as risk management and police resource allocation. The paper presents a comprehensive review of crime situation awareness and summarize the general process. Based on theories and algorithms, the review is categorized into methods based on criminal theories, machine learning, and deep learning. Additionally, it conducts an in-depth analysis of the limitations of existing methods and proposes three potential research directions: basic theories of crime situation awareness, crime situation awareness technology models, and crime situation awareness intelligence decision-making.

Keywords: Crime situation awareness; Criminal theory; Machine learning; Deep learning

1 INTRODUCTION

The modern social governance model should adhere to the principle of safety first and prevention first, establish a framework for comprehensive safety and emergency response, improve the public safety system, and promote the transformation of public safety governance model towards pre-emptive prevention. At present, the social security situation is still complex and severe, and various crimes such as burglary, telecommunications fraud, and online gambling continue to occur frequently, making the passive policing model based on investigation and crackdown difficult to meet the actual needs of public security work. Therefore, it is of great significance to rely on situational awareness technology to conduct in-depth mining of crime data, comprehensively grasp the public security situation, and enhance the ability to warn and prevent crime risks, maintain social stability, and realize the modernization of social governance capabilities.

Situation awareness technology is a method based on big data security, which enhances the ability to detect, understand, analyze, respond to, and handle security threats from a holistic perspective. Crime situational awareness technology is the scientific use of crime data to deeply analyze the spatial and temporal patterns and trends of crime events, predict possible future crime events, and provide decision support for public security work such as public security risk prevention and control, police resource scheduling, and so on. Early situational awareness methods mostly relied on specific criminological theories for analysis, such as the theory of daily activities [3], which defines crime as the result of the interaction of criminals, potential targets, and regulatory states in a specific space. The relevant methods are based on criminological theories, and include the construction of evaluation indicators and risk assessment models for potential offenders, community physical environments, and protective measures [4]. With the rapid development of information technology, the total amount of crime-related data is becoming increasingly large and the sources are more extensive, making high-performance data mining algorithms gradually become the mainstream of situational awareness technology [5]. For example, using historical crime data to train a crime situational awareness data model, and predicting future crime trends through existing data information and popularity; Through clustering analysis of crime data, potential similarities and correlations between different cases can be explored.

In summary, this article provides a comprehensive overview of crime situational awareness technology, with three contributions.

- Summarizing the general process of crime situational awareness technology.
- Propose a new classification method for crime situational awareness technology, which is divided into methods based on criminological theory, machine learning, and deep learning according to their principles.
- Conduct a systematic analysis and review of existing crime situational awareness technologies to understand the principles and applications of existing technologies.
- Prospect the future research direction of crime situational awareness technology.

2 CRIME AWARENESS PROCESS

Crime situation perception is the comprehensive perception and assessment of the overall situation of criminal activities through various data sources and analysis methods, in order to better carry out crime governance work. Typically, crime situational awareness generally includes the following steps:

2.1 Acquisition of Crime Data

Crime data sources are mainly divided into two aspects: (1) Internal data from police departments, such as police call records (including confidential information like case types, time of occurrence, location, participants, etc.) and urban surveillance videos; (2) Open-source external data, such as public information on government websites and user data on

social networking platforms. Since crime data often involves public safety and personal privacy issues, it is crucial to avoid illegal disclosure of relevant information during the acquisition process and to ensure proper data desensitization is carried out.

2.2 Crime Data Preprocessing

Crime data preprocessing involves converting the collected raw crime data into a format that is recognizable and storable by computers. This includes, for example, removing useless information such as data noise, correcting errors in characters and timestamps within the data, identifying and deleting duplicate records, and handling missing values. In addition, to improve data storage efficiency, data standardization operations are necessary, encompassing format unification (aligning information such as time and location), data encoding, data conversion, and data verification.

2.3 Crime Data Feature Engineering

Feature engineering involves transforming the preprocessed crime data by extracting and retaining specific information based on specific analysis tasks. For instance, the open-source data Chicago Crime Data [6] contains crime records from the Chicago Police Department's CLEAR system since 2001 (excluding murders), with 22 field features such as case number, date, latitude and longitude, crime type, community, arrest status, etc., to describe each crime event. Good feature engineering not only enhances the predictive performance of situational awareness technologies but also improves the interpretability of the perception results.

2.4 Analysis of Crime Data

Crime data analysis is a process of analyzing and interpreting crime-related data using criminological theories, data mining, machine learning, and other technical means. The core objectives are to identify crime patterns and perceive crime trends, assess the risk levels of crime events, determine the basic types of crime events, and predict potential criminal activities. By analyzing crime data, public security organs can more effectively allocate policing resources, formulate community and street security strategies, and enhance the level of urban public safety governance and urban resilience.

3 CRIME AWARENESS TECHNOLOGY

At present, research on crime situational awareness technology has made some progress, and many research results with great reference significance and high value have been formed. This section is divided into methods based on criminological theory, machine learning, and deep learning according to the theories and algorithms of different technology applications.

3.1 Method based on criminological theory

Classic criminological theories provide important theoretical support for situational awareness, such as the theory of daily activities, which suggests that crime is the result of the interaction of criminals, potential targets, and regulatory states in a specific space, and is associated with specific physical environments. The theory of crime patterns suggests that crime is most likely to occur in areas where the activity spaces of both perpetrators and victims overlap. Based on the above theory, Zhao Ziyu et al. [7] conducted an in-depth analysis of pickpocketing cases in Changchun City, and concluded that the spatial pattern of pickpocketing cases is proximity predation, which means that the high-incidence areas of pickpocketing and the main residential areas of criminals are highly overlapping in physical space, and there is a significant spatial attenuation effect. At the same time, based on existing case data, it is inferred that the crime buffer zone is roughly located 2 km away from the residence of the perpetrator. Zhao Pengkai and others used community property-related public security cases to reveal the risk sources of community property-related cases, and constructed evaluation indicators including potential perpetrators, potential victims, community physical environment, and protective measures. They determined the weight coefficients of indicators at all levels through the analytic hierarchy process [8], among which the high proportion of migrant population in the community, high coverage of technical defense, and high guard at community entrances and exits have a greater impact on property-related crimes. Based on the theory of environmental criminology and population and environmental factors, Chen Peng and others constructed a hierarchical model and evaluation index system for community burglary cases [9]. The weight of each indicator is shown in Table 1. From the results, the main factors affecting burglary cases are the proportion of migrant population and preventive measures. The method based on criminological theory relies on researchers to integrate theoretical knowledge of crime and typical cases, in order to discover the patterns and trends of crime.

3.2 Methods Based on Machine Learning

Research on machine learning-based methods explores how to use computers to obtain potential information contained in crime data, involving feature processing, situational awareness models, and other content. Sun Feifei and others extracted common characteristics of criminals in criminal cases from multiple dimensions such as life trajectory, growth experience, views on life, and ethical and moral concepts, forming a criminal personality profile. Then, they mapped the characteristics to a random forest to classify simulated criminal samples and perceive the criminal tendencies of key individuals [10]. Zhu Xiaobo et al. proposed the PSO-BP method to improve the sensitivity to initial weights and the tendency to fall into local optima in the analysis of crime data using traditional BP neural networks. By introducing the particle swarm optimization algorithm PSO to globally search and optimize network weights, the PSO-BP method can

effectively address these issues. In the experiment of predicting the number of theft crimes in Chicago, the PSO-BP neural network model significantly improved the prediction accuracy compared to the BP model, with the relative error reduced from 4.68% to 1.635%. Wang Juan and others have added analysis of environmental factors in crime areas, using ensemble learning methods [12] to mine the main environmental factors of crime and predict regional risk levels [13]. They have identified a temporal distribution pattern for robbery-related crimes, which occur most frequently around 6 pm and 10 pm in the evening, and more frequently in summer than in other seasons, with a certain periodicity. In addition, when Wang Juan and others conducted a classification prediction experiment on the criminal risk areas of robbery-related crimes, the ensemble learning model had the best classification effect, with an accuracy rate of over 90%.

Table 1 Risk Assessment Criteria Weight Allocation [9]

First-level indicators		Second-level indicators	
Personnel (B1)	0.4286	Proportion of floating population (C1)	0.3214
		Per capita income level (C2)	0.1071
		Community openness (C3)	0.0476
Environment (B2)	0.1429	Building security (C4)	0.0476
		Remoteness of community location (C5)	0.0476
		Coverage of technological prevention measures (C6)	0.1837
Prevention (B3)	0.4286	Proportion of human prevention forces (C7)	0.0612
		Propaganda intensity for community theft prevention (C8)	0.1837

3.3 Methods Based on Deep Learning

Deep learning is a new research direction in the field of machine learning, which can learn the inherent rules of sample data more deeply. Shen Hanlei and others used an improved recurrent neural network, Recurrent Neural Network, RNN - Long Short-Term Memory Network, LSTM model to construct a binary alarm data long short-term memory model BD-LSTM and a frequency statistics data long short-term memory model RD-LSTM to predict the probability and number of cases occurring in each region [14]. In practice, Shen Hanlei and others used the data of burglary cases in the 110 alarm data of WH City from 2015 to 2018 to train the BD-LSTM and RD-LSTM models, effectively predicting the occurrence probability and number of burglary cases in various areas of WH City, with high accuracy and stability. Wei Dong and his team used deep neural networks and Mnd-Knox algorithm to capture the density of crime behavior in the microscopic scale, and visualized the results as a crime hotspot information map to guide the allocation of police resources [15]. Zhai Shengchang and others proposed a BP neural network model ARIMA-GRU that comprehensively considers factors such as season, time correlation, spatial correlation, holidays, and temperature to address the difficulty of capturing complex features in crime data. On the real crime dataset of Vancouver, the SARIMA-GRU model can effectively capture the above-mentioned complex features of crime, improving the prediction accuracy of urban remote areas and low-risk areas with fewer crimes [16].

3.4 Discussion

The above research provides important theoretical basis and technical methods for the research of crime situational awareness technology. However, there are still some shortcomings. First, the method based on criminological theory relies on manual analysis and research of crime theories and typical cases. The method is simple and the number of indicators is small, and it lacks analysis of important influencing factors such as spatial and temporal correlation. At the same time, it is limited by the theoretical and operational abilities of staff, and has limited guidance for practical police work; Secondly, methods based on machine learning and deep learning fail to effectively capture the spatiotemporal dependencies of crime data. Most models represent spatial information as grid data and temporal information as sequence data, ignoring the distance information in spatial data and the contextual information of spatiotemporal data during computation; Thirdly, there is a lack of systematic and specialized crime situation perception theory, and a universal crime situation perception theory model has not yet been constructed. There is also a lack of targeted, standardized, and practical risk warning and prevention mechanisms.

4 APPLICATIONS OF CRIME SITUATION PERCEPTION

Crime situational awareness can improve the quality of crime monitoring and prediction through emerging technologies such as big data, artificial intelligence, and the Internet of Things, thereby promoting the construction of public security information and raising the level of public safety governance. In combination with the technologies related to crime situational awareness in Section 3, the application scenarios of crime situational awareness mainly include the prediction of crime hotspots, the prediction of crime frequency, and the prediction of accomplice relationships.

4.1 Prediction of a crime hotspot area

Crime hotspot prediction is a qualitative analysis technique aimed at predicting the spatial distribution of future crime events. First, divide the city map into several regions, then extract crime event features from each region, and then input

these features into machine learning algorithms such as kernel density function or random forest to infer whether the crime rate in each region is a hot spot. The key issues in predicting crime hotspots are the division of spatial regions and the extraction of crime characteristics. These operations directly affect the prediction granularity and accuracy of the model. Through the prediction of crime hotspots, it can not only provide a scientific basis for the public security organs to formulate prevention strategies, but also play an important role in the allocation of police resources, the optimization of patrol routes, and the formulation of community safety policies.

4.2 Prediction of crime frequency

The purpose of crime frequency prediction is to analyze and mine crime time series data through deep learning models such as RNN to predict the frequency of future crimes. By collecting and processing historical crime data, researchers train RNN models to learn and understand the temporal patterns, periodicity, and seasonal changes in crime occurrence, and analyze crime trends. In addition, the RNN-based crime frequency prediction method can effectively handle and capture the time-series dependencies in crime records, and retain the nonlinear features contained in the data, achieving high-precision crime frequency prediction and providing a scientific basis for resource allocation and prevention strategies for law enforcement departments.

4.3 Prediction of the relationship

Prediction of accomplice relationships is a method for identifying potential accomplice connections by analyzing the relational structure of users in social network data. This technique is based on social network analysis theory and utilizes indicators such as node centrality, connectivity, and clustering coefficient to reveal the roles of individuals within the criminal network and their associated patterns. Simultaneously, by constructing and analyzing social network graph structure data and inputting it into graph neural networks or community detection algorithms, it is possible to effectively predict accomplice relationships and identify core members of criminal gangs. In particular, accomplice relationship prediction based on deep learning not only aids in understanding the dynamic structure and operational mechanisms of criminal organizations during intelligence analysis, but also provides technical and decision-making support for public security intelligence analysis and research.

5 FUTURE RESEARCH DIRECTIONS

Crime situational awareness technology is a complex systems engineering that requires a scientific and reasonable architecture design. Although scholars have conducted research on situational awareness and public safety intelligence, a systematic theory has not yet been formed. Additionally, crime data exhibits characteristics such as dynamism, correlation, and uncertainty in both temporal and spatial dimensions, with extremely complex influencing factors, which significantly increases the difficulty of crime data mining and intelligence analysis. Based on this, this section proposes three potential research directions.

5.1 Establishing the Basic Theory of Crime Situational Awareness

Situational awareness technology has facilitated the transformation of crime governance models towards proactive prevention, shifting police work from a reactive mode focused on investigation and crackdown to a proactive mode centered on risk prevention and control. It enables the perception of the spatio-temporal patterns and trends of criminal events, thereby dissolving crime risks at their source. Therefore, it is crucial to clarify the concepts related to crime situational awareness, delineate the scope of research, review regulations, policies, and domestic and international research literature, summarize the progress, advantages, and disadvantages of existing work, analyze current challenges and development trends, and thereby lay a theoretical foundation for the subsequent construction of theoretical and technical models of crime situational awareness, as well as the application of situational awareness technology. Figure 1 summarizes the thinking and process of basic theoretical research on crime situational awareness.

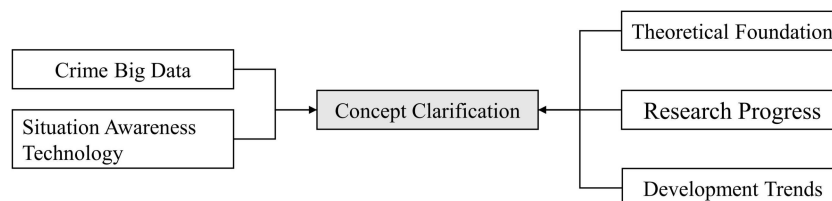


Figure 1 Basic Theories of Crime Situation Awareness

5.2 Enhancing Technical Models for Crime Situational Awareness

Crime data exhibits characteristics such as dynamism, correlation, and uncertainty in both temporal and spatial dimensions, and the factors influencing crime occurrence are also extremely complex, encompassing numerous aspects including population, education level, poverty rate, employment, climate, and more. Most existing situational awareness models represent spatial information as grid data and temporal information as sequence data, neglecting the distance information within spatial data and the contextual association information within spatio-temporal data during computations. To address these issues, we convert criminal factor characteristics, temporal data, and spatial data into spatio-temporal graph data, and construct a crime risk situational awareness model based on spatio-temporal graph representation learning [17]. This model extracts the spatio-temporal dependencies within crime data, improves the

accuracy of crime trend predictions, and generates crime hotspot information maps. Figure 2 summarizes the thinking and process of research on technical models for crime situational awareness.

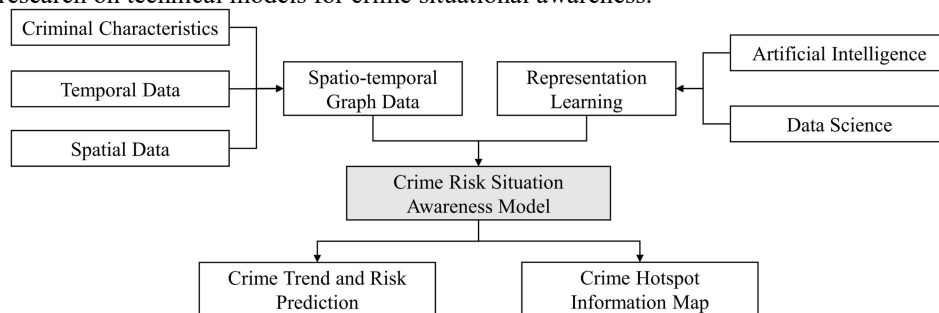


Figure 2 Technical Models of Crime Situation Awareness

5.3 Applying Crime Situational Awareness Technology for Scientific Decision-Making

The application of crime situational awareness technology is based on the fundamental connotation of situational awareness and guided by the results of situational awareness technical models. It tracks the development and evolution of crime risks and promptly pushes security intelligence to decision-makers, thereby enabling accurate and efficient intervention in public safety governance. Crime situational awareness is a complex systems engineering that primarily encompasses three aspects: risk monitoring, decision-making response, and post-event evaluation. Risk monitoring involves tracking the development trends of risks, promptly pushing intelligence to decision-makers, and intervening in public safety governance in a real-time, scientific, and efficient manner. Decision-making response involves adopting scientific and effective measures based on decision-making scenarios, risk evolution pathways, occurrence conditions, and the logical relationships among these conditions, to disrupt the evolution of security risks. Post-event evaluation relies on situational awareness to prevent the occurrence of secondary and derivative events, and carries out damage assessment, event traceback, intelligence evaluation, and other tasks. Figure 3 summarizes the thinking and process of applying crime situational awareness technology for intelligence decision-making.

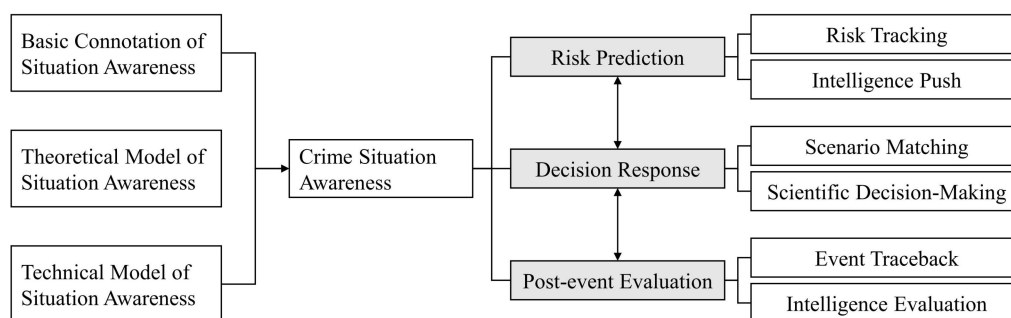


Figure 3 Intelligence Decision-Making for Crime Situation Awareness

6 CONCLUSION

This paper provides a comprehensive review of the literature related to crime situational awareness technology and proposes a classification method for situational awareness technologies. Meanwhile, it conducts an in-depth analysis of the shortcomings of existing methods and suggests three potential research directions. In the face of new social security challenges, public security organs must establish a new policing model supported by big data empowerment, characterized by "professionalism + mechanisms + big data." In future work, it is crucial to fully leverage crime situational awareness technology to facilitate the transition of policing from a reactive mode focused on investigation and crackdown to a proactive mode centered on risk prevention and control. Additionally, relevant models should be utilized to analyze the evolution pathways, occurrence conditions, and logical relationships among various crime risks, providing reference for relevant government decision-making departments and promoting the transformation of public safety governance towards proactive prevention.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

FUNDING

This work was supported in part by the Social Science Fund of Guangxi under Grant 23FTQ005 and the Young and Middle-aged Teacher's Research Ability Enhancement Project of Guangxi under Grant 2024KY0904.

REFERENCES

- [1] Guo Yu, Liu Fangyu, Zhang Chuanyang. Research on the event knowledge graph of public security event driven by multimodal data. *Library and Information service*, 2024, 68(24): 15-26. DOI:10.13266/j.issn.0252-3116.2024.24.002
- [2] Wang Bing, Zhou Jiasheng, Shi Zhiyong. A model for safety & security situational awareness and shaping empowered by digital intelligence. *Library Tribune*, 2024: 1-9.
- [3] Liu L, Sun Q Y, Xiao L Z, et al. The temporal influence difference of drug-related personnels' routine activity on the spatial pattern of theft. *Journal of Geo-information Science*, 2021, 23(12): 2187-2200.
- [4] Xu Wenwen, Liu Yiliang. A pyramid model of community resilience construction in a risk society. *China Safety Science Journal*, 2023, 33(9): 189-195.
- [5] He R X, Tang Z D, Jiang C, et al. A graph convolution-based spatio-temporal crime prediction model considering road weights. *Journal of Geo-information Science*, 2023, 25(10): 1986-1999.
- [6] Walter R J, Tillyer M S, Acolin A. Spatiotemporal crime patterns across six US cities: Analyzing stability and change in clusters and outliers. *Journal of Quantitative Criminology*, 2023, 39(4): 951-974.
- [7] Zhao Ziyu, Liu Daqian, Xiao Jianhong, Wang Shijun. Spatial characteristics and influencing factors analysis of journey-to-crime based on crime pattern theory: A study of theft crime in Nanguan District, Changchun. *Geographical Research*, 2021, 40(03): 885-899.
- [8] Zhao Pengkai, Chen Peng, Hong Weijun. Research on evaluation index system for property-related public security risks in communities based on analytic hierarchy process (AHP). *Legal System and Society*, 2015, (26): 173-175.
- [9] Chen Peng, Hu Xiaofeng, Zhang Chao. Research on risk evaluation model for burglary cases in communities. *Journal of Chinese People's Public Security University(Science and Technology)*, 2015, 21(02): 76-80.
- [10] Sun Feifei, Cao Zhuo, Xiao Xiaolei. Application of an improved random forest based classifier in crime prediction domain. *Journal of Intelligence*, 2014, 33(10): 148-152.
- [11] Zhu Xiaobo, Ci Jinfang. Application of improved pso-bp neural network algorithm in the prediction of theft crime. *Computer Applications and Software*, 2020, 37(01): 37-42+75.
- [12] Luo Changwei, Wang Shuangshuang, Yin Junsong, et al. Research Status and Prospect of Ensemble Learning. *Journal of Command and Control*, 2023, 9(01): 1-8.
- [13] Wang Juan, Long Junzhou, Guan Yuxiang. Analysis and prediction of robbery crimes based on stacking ensemble learning. *The Journal of Yunnan Police College*, 2023, (05): 114-123.
- [14] Shen Hanlei, Zhang Hu, Zhang Yaofeng, et al. Research on burglary crime prediction based on long short-term memory model. *Statistics & Information Forum* 2019, 34(11): 107-115.
- [15] Wei Dong, Zhang Tian-yi. Application of knox feature optimization in grid crime spatio-temporal prediction. *Journal of Chinese Computer Systems*, 2022, 43(11): 2456-2464.
- [16] Zhai Shengchang, Han Xiaohong, Wang Li, et al. SARIMA-GRU crime prediction model based on nonlinear Combination of BP neuralnetwork. *Journal of Taiyuan University of Technology*, 2023, 54(3): 525-533.
- [17] Zhao D, Du P, Liu T, et al. Spatio-temporal distribution prediction model of urban theft by fusing graph autoencoder and GRU. *Journal of Geo-information Science*, 2023, 25(7): 1448-1463
- [18] Hu K, Li L, Tao X, et al. Information fusion in crime event analysis: A decade survey on data, features and models. *Information Fusion*, 2023, 100: 101904.

A-SHARE INTELLIGENT STOCK SELECTION STRATEGY BASED ON THE DEEPSEEK LARGE MODEL: TECHNICAL ROUTES, FACTOR SYSTEMS, AND EMPIRICAL RESEARCH

HaiLong Liao

School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China.

Corresponding Email: jnhailong@126.com

Abstract: This research focuses on the application of the DeepSeek large - scale model in intelligent stock selection in the A - share market. It constructs a multi - dimensional factor analysis framework that integrates reinforcement learning and a mixture - of - experts architecture. Through empirical research, the advantages of this model in terms of return acquisition and risk control are verified, providing new intelligent strategies for A - share investment and promoting technological innovation and development in the capital market.

Keywords: DeepSeek large - scale model; Intelligent stock selection in the A - share market; Factor system; Empirical research

1 INTRODUCTION

The global financial market is in a period of rapid transformation, with increasingly prominent complexity and information explosion characteristics. Against this background, traditional quantitative models face numerous challenges in investment decision - making. For example, the Fama - French three - factor model [1] has an explanatory power of only 37.2% for the TMT sector, highlighting the insufficiency of traditional quantitative models in capturing non - linear correlations. According to a report by a foreign research institution cited in The New Yorker magazine, ChatGPT responds to approximately 200 million requests per day and consumes over 500,000 kWh of electricity in the process. That is to say, ChatGPT's daily electricity consumption is equivalent to that of 17,000 American households [2]. The high processing cost limits its application in high - frequency trading scenarios. In addition, the market changes dynamically and frequently. In 2023, the industry rotation frequency of the A - share market increased by 28% year - on - year. Traditional models lack an effective dynamic environment adaptation mechanism and are difficult to keep up with the pace of market changes, greatly reducing the timeliness and accuracy of investment decisions.

The DeepSeek technology system, with its architectural innovation (MLA multi - head latent attention mechanism) and breakthrough in training paradigms (GRPO reinforcement learning algorithm), brings new opportunities to solve the above - mentioned problems [3]. It shows unique advantages in computational efficiency, resource allocation, and decision - making path optimization, injecting new vitality into the financial investment field.

This research focuses on the application of the DeepSeek large - scale model[4] in intelligent stock selection in the A - share market. It constructs a multi - dimensional factor analysis framework that integrates reinforcement learning and a mixture - of - experts architecture. This research has multiple important values. On the one hand, it is the first to systematically verify the implementation effectiveness of Chinese open - source large models in complex financial scenarios, filling a research gap in the relevant field. By constructing an interpretable AI - enhanced factor system, it provides investors with a more scientific and comprehensive investment analysis perspective. On the other hand, it demonstrates the optimization effect of edge - side deployment on the investment decision - making chain. Experiments show that the inference latency of the 7B model is less than 50 ms, which is of great significance for improving the efficiency of investment decision - making. The results of this research provide empirical support for the popularization of AI technology, and are expected to promote the transformation of the Chinese capital market towards intelligence, helping small and medium - sized investors gain a more favorable investment position in a complex market environment.

2 HISTORICAL DATA AND TREND ANALYSIS OF THE A - SHARE MARKET

2.1 Review of Historical Data

Since its establishment, the A - share market has gone through multiple development stages. The market scale has been continuously expanding, the investor structure has been gradually optimized, and the market system has become increasingly perfect. Reviewing historical data can provide rich reference for the current research.

Looking at the overall market trend, the Shanghai Composite Index has shown obvious cyclical fluctuations over the past few decades. For example, during the period from 2005 to 2007, the Shanghai Composite Index soared from 998 points to 6124 points, with a significant bull market. This was mainly due to factors such as the institutional dividend released by the share - splitting reform, the high - speed growth of the macro - economy, and a large amount of capital flowing into the stock market. Subsequently, in 2008, affected by the global financial crisis, the Shanghai Composite

Index dropped sharply to 1664 points. Market panic spread, and investors' confidence was frustrated. In 2014 - 2015, there was another rapid upward trend. The Shanghai Composite Index rose from around 2000 points to 5178 points. This stage was closely related to factors such as loose monetary policy and the development of margin trading business. However, the market then adjusted rapidly, resulting in a stock market crash and causing huge losses to investors.

In terms of industry performance, the price changes of different industries vary significantly in different periods. In the early stage, traditional energy and financial industries dominated the market. With the adjustment of the economic structure and the development of technology, consumer and technology industries have gradually emerged. For example, during 2019 - 2020, the liquor industry benefited from consumption upgrading and increased brand concentration. The stock prices of leading enterprises such as Kweichow Moutai and Wuliangye increased significantly. The semiconductor industry became a market hotspot against the background of national policy support and strong domestic substitution demand. Enterprises such as SMIC and GigaDevice Semiconductor performed outstandingly.

2.2 Trend Analysis

In recent years, the A - share market has shown some new trends. First, the degree of market institutionalization has been continuously increasing. With the continuous inflow of long - term funds such as social security funds, pension funds, and foreign capital, as well as the continuous expansion of the scale of public and private funds, institutional investors have gradually strengthened their say in the market. This makes the market investment style pay more attention to fundamental analysis and long - term investment value, and the stock price trend becomes more rational.

Second, scientific and technological innovation has become the core driving force of the market. Under the background of the country's strong promotion of the innovation - driven development strategy, the technology industry has developed rapidly, and its weight in the A - share market has been continuously increasing. From 5G communication, artificial intelligence to new energy vehicles, related industrial chain enterprises have emerged in an endless stream, bringing new investment opportunities to the market. For example, the new energy vehicle industry, driven by multiple factors such as policy subsidies, technological progress, and growing market demand, has seen a significant increase in the market value of enterprises such as BYD and Contemporary Amperex Technology Co., Limited, making the entire industry sector the focus of the market.

Third, the correlation between the market and the macro - economy has become closer. Changes in macro - economic data, such as GDP growth rate, inflation rate, and monetary policy adjustments, have a more significant impact on the A - share market. In the economic recovery stage, the market often anticipates an increase in corporate earnings, and stock prices rise. When there is greater downward pressure on the economy, the market is more cautious, and stock prices fluctuate more violently.

3 DEEPSEEK TECHNICAL FOUNDATION AND FINANCIAL SCENARIO ADAPTABILITY

3.1 Analysis of Core Technologies

The technical advantages of DeepSeek are reflected in multiple dimensions, mainly including three aspects: computational efficiency innovation, dynamic resource allocation [5], and decision - making path optimization [6].

In terms of computational efficiency innovation, the MLA architecture is the key. When processing large - scale data, traditional Transformer models have high KV cache requirements, which limit their processing efficiency to a certain extent. The MLA architecture effectively reduces the KV cache requirements through implicit attention mapping. When processing input sequences of different lengths, the MLA technology can effectively reduce the cache size, thereby improving the performance of the model. For example, in DeepSeek - V3, after using the MLA technology, the compression ratio of the KV cache reached 6 times. This means that when processing text sequences of the same length, the MLA technology can significantly reduce the required memory space, thus improving the operation efficiency of the model [7]. This improvement enables DeepSeek to support the real - time processing of high - frequency market data. In the rapidly changing financial market, it can timely capture market information and provide timely data support for investment decisions.

The DeepSeekMoE mixture - of - experts model realizes dynamic resource allocation. The model adopts a dynamic parameter activation mechanism. DeepSeekMoE is an innovative mixture - of - experts (MoE) architecture, aiming to achieve higher expert specialization and computational efficiency through fine - grained expert segmentation and shared expert isolation strategies. DeepSeekMoE further divides experts, enabling each expert to focus more on specific knowledge fields or tasks. This fine - grained division allows the model to improve the effect and efficiency by flexibly combining multiple experts when dealing with complex tasks. For example, in the DeepSeekMoE 16B model, 8 experts are selected from 64 experts for activation, thereby achieving higher knowledge acquisition accuracy and computational efficiency [8]. This mechanism, while ensuring the model capacity, significantly reduces the inference energy consumption by 70%. In financial scenarios, a large number of computational tasks require huge computational resources. This feature of the DeepSeekMoE model enables more efficient completion of complex financial analysis tasks under limited computational resource conditions.

Decision - making path optimization is another important advantage of DeepSeek. The GRPO (Generalized Reward Policy Optimization) algorithm allows the model to autonomously optimize investment strategies in an unsupervised environment by designing a multi - objective reward function [9]. In the backtest of the R1 version, its self - correction accuracy rate is as high as 82.3%. This means that the model can continuously adjust investment strategies according to

market changes, improving the accuracy and adaptability of investment decisions. For example, when the market fluctuates, the model can adjust the position portfolio in a timely manner according to the feedback of the reward function, reducing risks and increasing returns.

3.2 Adaptability to Financial Scenarios

In the application of the A - share market, DeepSeek shows three characteristics: long - text parsing, real - time response ability, and a localized knowledge base.

Long - text parsing ability is crucial for financial investment. In the financial field, annual reports and prospectuses contain a large amount of key information. Accurately extracting this information is of great significance for investment analysis. DeepSeek performs well in this regard. Its accuracy rate for extracting key information from annual reports/prospectuses has increased to 89.7%, which is significantly higher than 86.2% of GPT - 4, and the processing speed has increased by 3.2 times. This enables investors to obtain the core information of enterprises more quickly and accurately, providing a strong basis for investment decisions.

Real - time response ability is another outstanding advantage of DeepSeek in financial scenarios. The distilled 7B model can perform 1200 factor calculations per second on a device with 12GB of video memory. This calculation speed can meet the strict real - time requirements of intraday trading. In intraday trading, the market changes rapidly. Timely factor calculation and investment decisions can seize fleeting investment opportunities and increase investment returns. DeepSeek has also constructed a localized knowledge base. In response to the unique policy orientation of the A - share market, such as the screening of "specialized, refined, characteristic, and new" enterprises, a special fine - tuning data set has been constructed. This has increased the strategy specificity by 41.6%, enabling it to better adapt to the characteristics of the Chinese capital market and providing strong support for investors to explore high - quality investment targets that meet the policy orientation.

4 METHODOLOGY FOR CONSTRUCTING INTELLIGENT STOCK SELECTION MODELS

4.1 Data Layer: Multimodal Factor System

This research constructs a dynamic database containing 6 categories and 32 factors. These factors cover multiple dimensions such as value, growth, momentum, sentiment, industry chain, and policy, comprehensively reflecting the comprehensive situation of enterprises and the market environment.

Table 1 Multidimensional Indicators and AI Enhancement Methods for Each Factor Type

Factor Type	Typical Indicators	AI Enhancement Method
Value Factor	PE/PB/Dividend Yield	Industry - relative Valuation Deviation Correction
Growth Factor	ROE Growth Rate/R&D Expense Ratio	Patent Text Technical Barrier Assessment
Momentum Factor	20 - Day Volatility/Turnover Rate	Main Capital Flow Map Analysis
Sentiment Factor	Stock Forum Public Opinion Sentiment Value	R1 Model Inference Chain Confidence Weighting
Industry Chain Factor	Difference in Inventory Turnover Rates of Upstream and Downstream	V3 Code Parsing of Industry Database
Policy Factor	Degree of Matching with the "14th Five - Year Plan"	Semantic Embedding of the Government Work Report

In terms of value factors, typical indicators such as PE/PB and dividend yield are selected, and AI enhancement is carried out through industry - relative valuation deviation correction. When evaluating the value of an enterprise, although traditional PE/PB indicators are commonly used, they are easily affected by the overall industry valuation level. Through industry - relative valuation deviation correction, it is possible to more accurately judge the value position of an enterprise within the industry and avoid valuation misjudgments caused by industry - wide systematic factors.

Growth factors include indicators such as ROE growth rate and R&D expense ratio, and AI enhancement is carried out using patent text technical barrier assessment. R&D investment is an important driving force for the future development of an enterprise, and patents are an important manifestation of an enterprise's technical strength. By analyzing patent texts and evaluating the technical barriers of enterprises, it is possible to gain a deeper understanding of the growth potential of enterprises and provide a more comprehensive perspective for evaluating growth factors.

Typical indicators of momentum factors include 20 - day volatility and turnover rate, and AI enhancement is carried out with the help of main capital flow map analysis. The momentum effect of the market is of great significance in investment decision - making. The flow direction of main capital often indicates the short - term trend of the market. Through main capital flow map analysis, it is possible to more accurately grasp the market momentum and adjust investment strategies in a timely manner.

The sentiment factor takes the stock forum public opinion sentiment value as an indicator and performs AI enhancement through R1 model inference chain confidence weighting. The public opinion in stock forums reflects the emotions and

expectations of market participants. Although it contains a large amount of noise information, through the processing of the R1 model and confidence weighting, valuable sentiment signals can be extracted to provide a reference for investment decisions.

The industry chain factor selects the difference in inventory turnover rates of upstream and downstream as an indicator and uses V3 code parsing of the industry database for AI enhancement. In the industry chain, changes in the inventory turnover rates of upstream and downstream enterprises reflect the supply - demand relationship and operation efficiency of the industry chain. Through V3 code parsing of the industry database, it is possible to conduct a more in - depth analysis of industry chain factors and explore investment opportunities on the industry chain.

The policy factor takes the degree of matching with the "14th Five - Year Plan" as an indicator and performs AI enhancement through semantic embedding of the government work report. Policies have a profound impact on the capital market. The "14th Five - Year Plan" clarifies the country's development strategy and key support areas. Through semantic embedding of the government work report, it is possible to more accurately evaluate the fit between enterprises and policies and discover policy - driven investment opportunities.

4.2 Model Layer: Three - Stage Training Framework

The model layer adopts a three - stage training framework, including supervised fine - tuning (SFT), reinforcement learning with human feedback (RLHF), and adversarial training. Each stage plays a key role in improving the performance of the model.

In the supervised fine - tuning (SFT) stage, historical data of the TOP50 portfolio from 2018 to 2022 (including 30 basic factors) is input, and a preliminary stock selection probability distribution is output. At this time, the Hit@10 accuracy rate reaches 68.4%. In this stage, historical data is used to train the model, allowing the model to learn basic investment rules and data characteristics, laying a foundation for subsequent training.

In the reinforcement learning with human feedback (RLHF) stage, a reasonable reward function is designed: Sharpe ratio - maximum drawdown - industry diversification. The Sharpe ratio reflects the risk - adjusted return of an investment portfolio, the maximum drawdown reflects the maximum loss that an investment portfolio may face, and the industry diversification measures the distribution of the investment portfolio in different industries to avoid risks caused by over - concentrated investment. The Monte Carlo tree search is used to explore the position portfolio space, and the GRPO algorithm plays an important role in this process, reducing the annualized return volatility by 22%. In this stage, the model continuously tries different investment strategies in a simulated market environment and optimizes the strategies according to the feedback of the reward function, improving the model's investment decision - making ability.

In the adversarial training stage, in order to improve the robustness of the model, 10% noise data (such as financial fraud features, abnormal trading volume patterns) is injected into the data, and a discriminator network is constructed. Through adversarial training, the overfitting rate of the model is reduced from 17.3% to 5.1%, effectively enhancing the stability and reliability of the model when facing complex market environments and abnormal data.

4.3 Deployment Layer: Cloud - Edge - End Collaborative Architecture

The deployment layer adopts a cloud - edge - end collaborative architecture, giving full play to the advantages of different levels to achieve efficient investment decision - making support.

In terms of cloud - based training, the 70B model updates the industry knowledge base monthly, and this process consumes approximately \$1200 in computing power costs. The cloud has powerful computing resources and can support large - scale model training and data processing. By updating the industry knowledge base monthly, the model can timely obtain the latest industry information and market dynamics, maintaining sensitivity to market changes.

Edge inference deploys the 7B model at the brokerage trading terminal, with a latency of less than 100 ms. The trading terminal has extremely high requirements for real - time performance. The low - latency feature of edge inference enables investors to obtain the analysis results of the model in a timely manner during the trading process and make quick decisions.

In order to better adapt to market changes, a dynamic weight mechanism is also adopted, which automatically adjusts the factor weights according to market volatility. For example, during high - volatility periods, the weight of the momentum factor is increased by 15%. Changes in market volatility reflect the degree of market risk and uncertainty. By dynamically adjusting factor weights, it is possible to make more rational use of various factors in different market environments, improving the adaptability and effectiveness of investment strategies.

5 Empirical Analysis: Strategy Backtesting and Market Impact

5.1 Backtesting Results (January 2023 - June 2024)

Backtesting was conducted on the DeepSeek strategy, traditional multi - factor models, and the CSI 300 Index. The results are presented in the following table:

Table 2 Backtesting Data of Various Indicators under Different Strategies and Models

Indicator	DeepSeek Strategy	Traditional Multi - factor Model	CSI 300 Index
-----------	-------------------	----------------------------------	---------------

Indicator	DeepSeek Strategy	Traditional Multi - factor Model	CSI 300 Index
Annualized Return	32.1%	18.5%	6.7%
Sharpe Ratio	1.87	0.92	0.35
Maximum Drawdown	- 12.3%	- 22.7%	- 28.4%
Industry Coverage	82%	65%	100%

In terms of the annualized return, the DeepSeek strategy reached 32.1%, which is significantly higher than 18.5% of the traditional multi - factor model and 6.7% of the CSI 300 Index, indicating that this strategy has an obvious advantage in return acquisition. The Sharpe ratio measures the additional return that an investment portfolio can obtain over the risk - free return when taking on a unit of risk. The Sharpe ratio of the DeepSeek strategy is 1.87, much higher than 0.92 of the traditional multi - factor model and 0.35 of the CSI 300 Index, suggesting that it performs outstandingly in risk - adjusted returns. Regarding the maximum drawdown, the DeepSeek strategy is - 12.3%, which is a significant reduction compared to - 22.7% of the traditional multi - factor model and - 28.4% of the CSI 300 Index, reflecting the advantage of this strategy in risk control. In terms of industry coverage, the DeepSeek strategy reaches 82%, which is higher than 65% of the traditional multi - factor model. Although it is lower than the CSI 300 Index, it still indicates that this strategy can cover a relatively large number of industries and achieve a relatively wide investment layout.

5.2 Market Structure Impact

The DeepSeek strategy has had a variety of positive impacts on the market structure.

In terms of empowering small and medium - sized investors, after a private equity fund adopted the 7B edge - side model, the strategy research and development cost decreased from 2 million yuan per year to 450,000 yuan. This enables small and medium - sized investors to obtain advanced investment strategies at a relatively low cost, reducing the investment threshold, enhancing the competitiveness of small and medium - sized investors in the market, and promoting market fairness.

In terms of capturing industry rotation, during the artificial intelligence sector market in Q4 2023, the model identified targets such as Cambricon and Sugon two weeks in advance, and the portfolio's excess return reached 18.9%. The accurate industry rotation capture ability helps investors to promptly grasp market hotspots, adjust investment portfolios, and obtain higher returns.

In terms of improving liquidity, a 1 - basis - point increase in the trading volume share of the strategy can reduce the liquidity premium of the ChiNext Index by 0.3%. This indicates that the application of this strategy helps to improve market liquidity, reduce transaction costs, and enhance market operation efficiency.

6 CASE STUDIES: INVESTMENT OPPORTUNITIES DRIVEN BY DEEPSEEK

6.1 Computing Power Infrastructure

Taking Cambricon (688256) as an example, DeepSeek - V3 analyzed the performance parameters of its Siyuan 590 chip and predicted that the market share of AI chips in 2024 would increase to 19% (7 percentage points higher than the consensus forecast of securities firms). By comparing the supply chain data of NVIDIA H20, it was found that Cambricon had an advantage in packaging and testing costs. This analysis result provides investors with a new perspective on the investment value of Cambricon. Based on this, investors can more accurately evaluate the future development potential and investment returns of Cambricon, and thus make more rational investment decisions.

6.2 Financial Technology Applications

In the case of Hundsun Technologies (600570), the 7B model conducted real - time analysis of the customer feedback of its O45 system and captured the key signal that "the customized demand of asset management institutions increased by 23%". Combined with the policy deduction ability of the R1 model, it predicted an incremental market of 1.4 billion yuan brought by the new asset management regulations. This case demonstrates the application value of the DeepSeek model in the financial technology field. Through real - time analysis of customer feedback and in - depth interpretation of policies, potential investment opportunities can be explored, providing strong support for investors' investments in the financial technology sector.

6.3 Edge - side Hardware Ecosystem

By analyzing the technology roadmap of the robotics industry, the model found that the RK3588S chip of Rockchip (603893) accounted for 18% of the BOM cost of service robots. Combined with the V3 code generation ability, it simulated the performance elasticity brought by the mass production of humanoid robots. This analysis helps investors understand the important position of Rockchip in the edge - side hardware ecosystem and its potential performance growth space, providing an important basis for investing in Rockchip.

7 Challenges and Future Directions

7.1 Existing Challenges

In terms of regulatory compliance, the black - box decision - making process of the DeepSeek model conflicts with the "Algorithmic Management Guidelines for the Securities and Futures Industry" [10]. Since the decision - making process of the model is difficult to explain intuitively, regulatory authorities face difficulties in supervising investment strategies. To solve this problem, it is necessary to develop a SHAP value visualization interpretation module to present the decision - making logic of the model in a visual way, improve the transparency of the decision - making process, and meet regulatory requirements.

Data barriers are also an important issue. The acquisition cost of unstructured data accounts for 63% of the total model cost. Unstructured data, such as news reports and social media information, although containing rich market information, is difficult to acquire and process. Currently, it is necessary to explore federated learning solutions to carry out data collaboration and model training without disclosing the original data, reducing the data acquisition cost while protecting data privacy.

7.2 Evolution Path

In the future, the DeepSeek model has broad prospects for development in the field of intelligent stock selection in the A-share market. Its evolution path will revolve around three key directions: multimodal integration, real-time leapfrogging, and compliance innovation. By doing so, it will continuously enhance the model's performance and market adaptability, and further promote the intelligent transformation of A-share investment.

In terms of multimodal integration, there is great potential in accessing the DeepSeek-V3 visual module to analyze the industrial chain map. Take the photovoltaic industry as an example. By obtaining the workshop monitoring data of photovoltaic enterprises, the model can intuitively understand the operating status of enterprise production equipment, the efficiency of the production process, and the real-time situation of product quality.

Real-time leapfrogging is an important direction for the development of the DeepSeek model. The adoption of MTP (Multi-Token Prediction) technology can significantly improve the generation speed of trading signals, making it reach the microsecond level. In the current rapidly changing financial market, trading opportunities are fleeting. For example, in high-frequency trading scenarios, when there is sudden news or abnormal market fluctuations, the microsecond-level trading signal generation speed allows investors to react quickly.

Compliance innovation is a crucial aspect that must be emphasized in the development process of the DeepSeek model. Developing a "regulatory sandbox" version and embedding intelligent contracts for investor suitability management are important measures to achieve compliant development. The "regulatory sandbox" provides a safe testing environment for new investment strategies and models. In this environment, the model can operate under simulated market conditions while being monitored and evaluated in real time by regulatory authorities.

With the continuous deepening of multimodal integration, the gradual realization of real-time leapfrogging, and the continuous advancement of compliance innovation, the DeepSeek model will play an even more powerful role in the field of intelligent stock selection in the A-share market. It can not only provide investors with more accurate and efficient investment decision-making support, but also promote the intelligent process of the entire capital market, facilitate the optimal allocation of financial resources, and help the Chinese capital market move towards a stage of higher-quality development.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

- [1] Eugene F, Fama, Kenneth R, French. Fama - French Three - Factor Model and Its Application. 1993. Chinese Software Developer Network. Retrieved from <https://blog.csdn.net/hzk427/article/details/104187847>
- [2] Liu Yaning. ChatGPT Consumes Over 500,000 kWh of Electricity per Day. How Energy - Intensive Are AI Large - Scale Models? Titanium Media. 2024. Retrieved from <https://baijiahao.baidu.com/s?id=1793597173423030217&wfr=spider&for=pc>
- [3] Qiao Nan. Decoding the Innovation Path of DeepSeek: The Evolution Roadmap of Three Generations of Models. 2025. Retrieved from <https://news.sina.cn/ai/2025-02-09/detail-ineiwqye4694847.d.html>
- [4] Liao Hailong. DeepSeek Large - Scale Model: Technical Analysis and Development Prospect. Journal of Computer Science and Electrical Engineering, 2025, 7(1): 33-37. DOI: <https://doi.org/10.61784/jcsee3035>
- [5] Lieyan Benniu. DeepSeek: Algorithm Innovation Leads to an AI Computing Power Boom and Usher in a New Era of Intelligent Technology. 2025. Retrieved from http://mp.weixin.qq.com/s?__biz=Mzg2MTE3NDA2Mg==&mid=2247494898&idx=1&sn=b99504080dcd03e5a6809136486d5bde&scene=0
- [6] Wang Houdong. The Important Impact of DeepSeek on the Global Development of Artificial Intelligence. 2025. Retrieved from http://mp.weixin.qq.com/s?__biz=MjM5MjgzMzg1Nw==&mid=2453991839&idx=1&sn=74a9185fb47c86c8deb4939df9b81cd5&scene=0
- [7] Demon Akana. Introduction to the Technical Principles of Deepseek's MLA Technology. Chinese Software Developer Network. 2025. Retrieved from <https://blog.csdn.net/bestpasu/article/details/145539423>

- [8] Baitai Laoren. DeepSeekMoE Architecture. Chinese Software Developer Network. 2024. Retrieved from https://blog.csdn.net/weixin_41429382/article/details/144701990
- [9] Zhihong Shao, Peiyi Wang, Qihao Zhu, et al. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. 2024. DOI: <https://doi.org/10.48550/arXiv.2402.03300>. Retrieved from <https://arxiv.org/pdf/2402.03300>
- [10] Le Tutu. In - Depth Analysis of Other Potential Defects of DeepSeek. Eastmoney Wealth Channel. 2025. Retrieved from <https://caifuhao.eastmoney.com/news/20250202221538225857610>

CUSTOMER SEGMENTATION AND CHURN PREDICTION BASED ON K-MEANS AND RANDOM FOREST: A CASE STUDY OF E-COMMERCE DATA

ZhuoRan Li

School of Economics, Nanjing University of Finance & Economics, Nanjing 210023, Jiangsu, China.

Corresponding Email: wxfcg2021@126.com

Abstract: This study aims to segment customers using the application of the K-means clustering algorithm and predict customer churn using the random forest method. Transactional data were used, including the order date, customer name, region, logistics company, quantity bought, payment amount, and frequency bought. K-means clustering was applied to group customers into segments, while a random forest model was constructed to predict customer churn. K-means clustering could determine four customer segments with different purchasing habits. Random forest model could predict customer churn and could find that attributes such as payment value and region were the most significant to use while determining the probability of churn. Results of this study verify that employing K-means clustering and random forest simultaneously for customer segmentation and customer churn prediction is efficient and assists in obtaining considerable insights for precision marketing.

Keywords: K-means; Random forest; Customer segmentation; Churn prediction

1 INTRODUCTION

In the current era of digital economy, the expansion of e-commerce has revolutionized the competitive landscape significantly, compelling companies to adopt more advanced strategies in a bid to retain customers and facilitate sustainable growth. E-commerce websites have been accumulating massive amounts of customer data, including purchasing behavior, browsing history, and demographic information. These data contain valuable insights into customer behavior and preference that can lead businesses to effective personalized marketing and service optimization. Two significant activities among customer data analysis applications are customer segmentation and churn prediction. Effective customer segmentation enables businesses to personalize marketing to specific customer segments, thereby optimizing customer satisfaction and loyalty. Conversely, churn prediction can identify the potentially churned customers and help businesses take proactive retention measures. These two processes are the crux of customer relationship management (CRM) for e-commerce. Customer segmentation and churn prediction have attracted widespread attention from academia and the business world. Researchers have worked quite a lot in order to improve and develop methods for these tasks, with a lot of analysis achieved through the help of various data mining techniques, some of which are quite advanced, and also various methods [1,2].

1.1 Existing Research on Customer Segmentation

The process of customer segmentation is about splitting groups of customers into different kinds of subgroups, which are based on a variety of shared features or characteristics [3]. These can include things like demographics, behavior patterns, and individual preferences [4]. In the past, traditional ways of segmentation often depended on using simple demographic data. However, more recently, with the progress of machine learning, there have been techniques introduced that are much more complex. For instance, K-means clustering is commonly used because it is relatively simple and also effective in managing large sets of data, which makes it a good choice for many situations. Some studies showed how K-means could be applied to segment customers based on their buying habits and the brands they prefer [5]. This kind of segmentation could lead to insights that are very useful, particularly for marketing that is more targeted. There was also a comparison made between K-means and some other clustering methods, such as DBSCAN and agglomerative clustering [6]. In this comparison, it was found that K-means was generally the best performer in terms of two specific measures: the silhouette score and Davies-Bouldin score, which are used to evaluate clustering quality. Moreover, some of the studies that have been done have also explored how segmentation techniques can be connected with business strategies, in order to design marketing campaigns that are better suited to the specific needs and characteristics of each of the segmented customer groups [7].

1.2 Existing Research on Customer Churn Prediction

Customer churn prediction is something that looks at the people who might want to leave or end their connection with the business. It's important for the business because it affects their profits, so a lot of attention has been given to this. Many researchers have tried different ways, using various algorithms, and have focused on how selecting the right features and optimizing models can help to make predictions more accurate. A lot of churn prediction relies on machine learning methods, especially random forests, which are popular because they can handle large sets of data and are quite

strong in predicting churn. For example, one study from Deng and Gao made an improved K-means algorithm to split customers into groups, then used random forests to predict who might leave, and they were able to spot churners. Some other research also compared random forests with older algorithms, like decision trees or support vector machines, and found that random forests did a better job at predicting customer churn, outperforming them by a good margin. This shows that random forests are preferred in many cases when trying to figure out who's likely to leave a business.

1.3 Significance of Applying K-means and Random Forest in E-commerce Customer Segmentation and Churn Prediction

Despite there being progress in both of these fields, research that connects customer segmentation with churn prediction is still, in a way, exploratory. This is especially true in e-commerce. Most studies up to now have been treating these two areas as separate entities, not really taking advantage of the opportunities that might exist if they were to be combined. In the e-commerce industry, combining K-means clustering with random forest algorithms could have some benefits, multiple benefits, actually. For instance, K-means clustering helps businesses identify customer groups based on things like shopping habits, demographics, and other characteristics. This, in turn, may allow for some personalized marketing strategies, or it might help in creating them. Meanwhile, random forests give a strong structure for predicting if a customer will churn or not by figuring out which factors influence their loyalty and whether they'll stay with the brand. The combined use of these two methods can, possibly, improve how well businesses are able to segment customers and predict their likelihood to churn, thus leading to better ways to optimize strategies around retaining customers. This study will, or aims to, explore how both K-means clustering and random forest algorithms might be applied together for customer segmentation and churn prediction in e-commerce. It hopes that by doing this, we will gain a fuller understanding of customer behavior. This, ultimately, could support e-commerce businesses in developing more efficient strategies for managing relationships with their customers.

2 METHODS

2.1 Data Collection

The data set utilized in this study includes transaction records on an online shopping website between September 2023 and November 2023, with variables including order date, customer name, region, delivery company, purchase quantity, payment amount, and purchase frequency. Following data cleansing and imputation of missing values, a total of 6,419 records were realized over this timeframe.

2.2 Data Preprocessing

Preprocessing is the basic step of any data mining process. The data were cleaned to deal with missing values and outliers. The missing values were replaced with mean for numeric attributes and mode (most frequent value) for categorical attributes. The outliers were detected and dropped using the Z-score method. Normalization on the dataset was used so that all features will be contributing equally towards clustering.

2.3 Customer Segmentation Using K-means Clustering

K-means clustering, which is a popular unsupervised learning technique, can be used to divide data into K different groups or clusters [8]. The idea is that data points that fall into the same cluster are supposed to be quite similar to each other, whereas those in different clusters are not so much alike [9].

The process of executing this algorithm seems to be something like intuitive. First, it becomes important, or necessary, to determine the number of clusters, K. This step, the selection of K, is regarded as a rather critical one. Usually, in practice, the elbow method is often chosen to decide the optimal K number. The idea behind this elbow method is, or can be, that the sum of squared errors (SSE) is calculated for different values of K, which involves summing the squared distances from each data point to its corresponding cluster center. It is said that when K is small, and as the value of K increases, the SSE decreases considerably, because having more clusters seems to allow for better fitting of data points. However, if you continue increasing K, at a certain point, the decrease in SSE becomes not so sharp or might even level off. This, at some point, may result in a turning point, resembling an elbow, in the graph that shows the K value along with the SSE on the other axis. This turning point in the graph shows the optimal K value, which is believed to be the best.

Once K is chosen, K data points get selected randomly from the dataset, which may serve as the initial cluster centers for the process that will follow. After this selection, an iterative process begins, where in each iteration, the distances from all data points to the K cluster centers are calculated. Usually, a metric like Euclidean distance is employed to determine how far each data point is from a cluster center. Afterward, based on the calculated distance, each data point is assigned to the closest cluster center, or to the cluster that seems nearest. Once all the points have been assigned, the next step is to compute the mean value of the data points in each cluster, and then this mean serves to update the cluster center position. These steps of calculation and update continue to repeat, with the iterations taking place over and over again, until the cluster centers stop moving much or when the number of iterations reaches a preset limit. When the process reaches this point, the algorithm has reached convergence, meaning that the clustering result is obtained. In this

study, the elbow method is indeed used to figure out the best K number, with clustering depending on variables that include purchases, frequency, and amount.

2.4 Churn Prediction Based on Random Forest

Random forest is an ensemble decision tree learning algorithm and is typically applied in the field of machine learning. Random forest is especially popular due to its high predictive performance as well as robust performance.

To the algorithm principle, random forest creates numerous decision trees and aggregates the prediction results of the decision trees to obtain the final prediction conclusion [10]. In generating each decision tree, random forest adopts a double randomization process. On one hand, through the method of resampling with replacement from the initial training set, a number of bootstrap samples of equal size to the initial set are constructed [11]. Hence, each bootstrap sample may consist of duplicated data, and about 30% of data in the initial set will never appear in the bootstrap sample. Such data are known as out-of-bag data and can be used for model evaluation. On the other hand, when all nodes in the decision tree are split, not all features are utilized. Instead, a random subset of features is sampled from all features, and the best splitting feature is selected from this subset. This random process provides each decision tree some level of distinctiveness, and hence enhances the ability of the model to generalize.

After the model is constructed, for classification problems, random forest uses the voting approach, i.e., each decision tree predicts the sample by prediction, and finally, the category with the most votes is the prediction result of the random forest; for regression problems, the average method is used, and the prediction values of each decision tree are averaged to obtain the final prediction value.

In this paper, whether the customers purchasing this month will continue buying next month or not is shown as "churn" or "not churn". Customers purchasing next month are shown as "not churn", otherwise as "churn". "Whether churn" is used as the target variable, and other variables (order date, customer name, region, logistics company, payment amount, etc.) are used as input variables [8]. Non-numeric variables (such as order date, customer name, area, logistics company, etc.) are coded before they can be analyzed. Random forest modeling is used to determine the major causal factors of e-commerce customer churn in order to facilitate the prediction and prevention of customer churn.

2.5 Model Integration

The model proposed, it integrates two methods, K-means clustering and random forest. The purpose is to better understand customer behavior, and it does this by splitting customers into different segments through clustering. Then, using random forest, predictions about customer churn are made. The combination of these two methods, in theory, it works by taking advantage of the strengths from both of them. This can give insights that are useful for businesses, like those in e-commerce, to target customers that belong to particular segments.

2.6 Experimental Setup

In this experiment, the tool used is Python, and it includes libraries like scikit-learn and pandas. The dataset for the random forest model, it is divided into two parts: a training set and a test set. The training set gets 70% of the data, and the test set gets 30%. When training, four-fold cross-validation is also applied. This training happens using 70% of the data, and then the remaining 30% is tested. This method, it ensures that the model is evaluated on unseen data after being trained on a large enough portion.

3 RESULTS

3.1 Customer Segmentation Based on K-means Clustering

Table 1 Customer Segmentation

Cluster types	center value		
	Purchase quantity	Payment amount	Purchase frequency
1	5.09	227.73	1.01
2	49	11298.195	49
3	17.67	3865.65	17.67
4	5.17	1208.80	5.17

For K-means clustering, the customers were grouped into four segments. Segment 1 is of consistent purchasing behavior but with low purchase frequency but relatively high payment amount per purchase. They must have special demand for the products but purchase with low frequency. Their purchasing behavior is rational, and they expect a specific quality of the product and services. These can be approached by the firms with high-cost-performance offerings or bundles of services so that they buy more often and remain loyal. Segment 2 includes high-value customers with high frequency of purchase and high payment amount. They are strongly brand-loyal and are the most important customer base for the company. Companies must prioritize retaining these customers by providing quality after-sales service and

personalized attention to enhance their word-of-mouth communication and brand loyalty. Segment 3's payment and purchase frequency are balanced, which means stable consumer behavior. These customers likely have stable demand for products. Companies can target this segment with promotional offers or membership schemes to increase their purchase frequency and spending. Segment 4 is defined by low value of payment and purchase frequency, which are low-expenditure buyers. They are price sensitive and require some type of focused marketing efforts that will improve their participation and consumption levels. Companies can provide low-price promotions or coupons to induce this segment to purchase (See Table 1).

3.2 Customer Churn Prediction using Random Forest

The impact of random forest on data classification is measured quantitatively by indicators. The accuracy rate and recall rate are better, the higher they are. During this test, the accuracy rate and recall rate of the cross-validation set and test set were checked simultaneously. It can be observed that the two indicators are relatively high. Since the precision rate and recall rate have an effect on one another, if balance between the two needs to be ensured, then the F1 - measure needs to be utilized. The results of the experiments show that the harmonic mean of the precision rate and recall rate of the F1 - measure is also fairly high. The smaller the AUC value, the better the effect of classification (See Table 2).

Table 2 Classification Evaluation Indicators for Training and Testing sets

	accuracy	recall	precision	F1	AUC
Cross validation set	0.944	0.944	0.959	0.944	0.884
Test set	0.9	0.9	0.911	0.886	0.956

Table 3 Feature Importance Score for Loss Prediction

Name	Importance
Freny	0.00%
Purchase quantity	1.30%
Actual payment amount by the buyer	78.10%
province	20.50%

The model indicates that payment amount is the most important attribute that affects customer churn, followed by customer location, and then purchase quantity and frequency of purchases (See Table 3). Why the payment amount is the most important attribute that affects customer churn is as follows. From the perspective of consumption ability and loyalty, the total payment value is the consumption ability and value contribution of the customer to the store. In general, high cumulative payment value customers indicate that they have high awareness of the store's products or services, have established some consumption habits and loyalty, and are more likely to continue creating value for the store. So, there is a very low possibility of their churn. On the other hand, customers with minimal aggregate payment may still be under the trial phase of the store, or lack too much familiarity with the products and services of the store, thus they will churn frequently. According to the input - output psychological view, from the customer's psychological stand, the more the customers spend within the store, the more they will experience that they spent considerable money and time costs in this store. To achieve the best return of the cost incurred by them, they will continue spending in this shop rather than easily moving to other shops.

The reason why customer location plays such an important role in customer churn is the following. From a perspective of geographical differences in consumption culture, the customers of this type of store in various geographical regions may also have different consumption habits and cultures. The customers in some areas can generally be more eager in demand and preferred for this type of store's products or services, and stronger in consumption desire. Thus, customers of such regions are stable, and churn is moderate. Customers in certain regions will have less demand for such kind of products or services, or other substitutes exist in the local market that are more competitive. Such customers will churn more frequently. From the perspective of logistics and convenience of service, the distance between the customer source area and the store and the convenience of logistics distribution will also affect the customer churn rate. Logistics speed is faster in close areas, customers get goods faster, and after-sales service is more convenient. The more convenient customer experience will encourage them to continue shopping at this store. In remote regions with inconvenient logistics, due to long delivery time, high freight costs, or easy occurrence of logistics problems, customer satisfaction may decrease, thus the probability of churn. From the perspective of regional differences in marketing impacts, the marketing investment of the store and the initial implementation of marketing measures in different regions are different. In a few of the key-marketed locations, the customers are more aware and store favorable, and it is easier to establish a stable customer base. In low-marketing coverage areas, the customers have less knowledge of the store

and tend to be impacted by other rival stores and defect.

As for the reason why purchase quantity and purchase frequency have relatively smaller effects on customer churn, it is because the purchase quantity and purchase frequency are subject to various factors, such as the special demand cycle of the customer and temporary changes in the external environment. Reasonably speaking, the amount bought and the buying frequency could only reflect the purchase frequency behavior and customer quantity within a specific time period, but cannot reflect fully the customer's loyalty to the store and long-term consumption intention. It is possible that customers will buy in bulk at a single instance due to incidental requirements or promotional events, or have high purchase frequency within a specified period of time, but this does not mean that they will continue buying. Therefore, purchase quantity and frequency shifts cannot accurately reflect whether customers will churn or not. When compared with the total payment amount and customer source region, their performance for predicting customer churn is rather poor.

4 DISCUSSION

The results of this study validate that the combination of the K - means clustering and random forest algorithms is effective in analyzing customer behavior and predicting customer churn. Customer segmentation using K - means clustering enables understanding in - depth of the purchasing habits and preferences of different customer segments. By customer segmentation using different customer segments, companies can develop focused marketing strategies to enhance customer satisfaction and loyalty. The high predictive capability of the random forest model of customer churn signifies the importance of key features such as payment value and customer source region. They have critical contributions towards customer loyalty and identifying prospective churners.

When examining the joint analysis outcomes of the two models, it can be found that customers with large cumulative payment amount during the three-month experiment period not only exhibit high purchase frequency and high purchase frequency but also share a relatively concentrated purchase region source. These customers are brand and store loyal and can provide high stability. By contrast, customers with a low purchase amount not only have a low number of purchases and low purchase frequency but also have a dispersed region source, low stability, and are likely to churn. Especially, low - spending customers and price - sensitive customers in the above clusters suffer from serious churn.

Based on this, companies can take a set of measures to increase sales and enhance customer stability, stickiness, and loyalty based on the combined analysis result of the two models. For customer segments with large payment volume and significant purchasing power, companies have to focus on retaining them by rewarding them and improving the shopping experience, e.g., embracing selective offers and personalized services to enhance their shopping pleasure. Especially for regions where there are a large number of customers, considering that they are most likely to be the source of big customers, it is better to increase the intensity of targeted development and marketing, and incline the promotion of direct - access traffic and advertising investment towards these regions. For price-conscious customers with low frequency of consumption, considering their low loyalty and simplicity of their transfer and churn due to small discounts such as price adjustments and their low requirements of brand and quality adjustments, it is recommended to increase their frequency of consumption and rate of purchase conversion through the use of discounts, promotion activities, and accurate recommendations.

In addition, this study also proposes the potential to explore additional characteristics and more advanced models to refine the accuracy of churn prediction. For example, the introduction of more advanced time-series data as well as the integration of variables like the time of last purchase would provide a greater understanding of customer behavior shift. Furthermore, exploring other machine-learning techniques such as gradient boosting or neural networks could deliver better performance for customer-churn prediction.

5 CONCLUSION

The study has been successful in utilizing the K - means clustering and random forest algorithm for churn prediction and customer segmentation in the e - commerce industry. The results highlight the importance of payment value and geographical distribution in churn prediction and provide insightful implications for targeted marketing and customer retention techniques. By identifying different customer segments and forecast customer churn well, firms are able to act proactively in addressing potential customer churn problems and enhance customer loyalty. The study confirms that combining K - means clustering and random forest is an efficient approach for customer segmentation and churn prediction and provides a workable model for big data analysis application in e - commerce sector. The findings of this study are also expected to add to the theoretical richness of customer behavior and provide actionable guidance to companies seeking to improve their data - driven decision - making.

Even though the results are encouraging, there are still some limitations to this study. The data for this study were collected from a particular e-commerce platform, and it may not be generalizable to other settings. The applicability of the analytical methods to other datasets and industries has yet to be further tested. In addition, this study did not take into account any external factors influencing customer behavior, such as market trends and economic conditions, that may potentially impact the results and need to be investigated further.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

- [1] Shweta Pandey, Neeraj Pandey, Deepak Chawla. Market segmentation based on customer experience dimensions extracted from online reviews using data mining. *Journal of Consumer Marketing*, 2023, 40(7): 854-868. DOI: <https://doi.org/10.1108/jcm-10-2022-5654>.
- [2] Petra Jilková. Customer Behaviour and B2C Client Segmentation in Data-Driven Society. *International Advances in Economic Research*, 2020, 26(3): 325-326. DOI: <https://doi.org/10.1007/s11294-020-09799-9>.
- [3] Tiffany S, Legendre. Consumer value-based edible insect market segmentation [edible insect market segmentation]. *Entomological Research*, 2020, 51(1): 55-61. DOI: <https://doi.org/10.1111/1748-5967.12490>.
- [4] Deepak Jaiswal, Vikrant Kaushal, Pankaj Singh, et al. Green market segmentation and consumer profiling: a cluster approach to an emerging consumer market. *Benchmarking: An International Journal*, 2020, 28(3): 792-812. DOI: <https://doi.org/10.1108/bij-05-2020-0247>.
- [5] Rui Zhao. CVM Model of Customer Purchasing Behavior Based on Clustering Analysis. *Proceedings of the 2021 3rd International Conference on Economic Management and Cultural Industry (ICEMCI 2021)*. 2021. DOI: <https://doi.org/10.2991/assehr.k.211209.328>.
- [6] Safae Bouhout, Youness Oubenaalla, El Habib Nfaoui. Comparative Study of Two Parallel Algorithm K-Means and DBSCAN Clustering on Spark Platform. *Advanced Intelligent Systems for Sustainable Development (AI2SD' 2020)*. AI2SD 2020. *Advances in Intelligent Systems and Computing*, 2022, 1418: 245-262. DOI: https://doi.org/10.1007/978-3-030-90639-9_20.
- [7] Wolfgang Bellotti, Daniela N. Davies, Y H Wang. Improved Multi-index Customer Segmentation Model Research. *International journal of smart business and technology*, 2021, 9 (2): 49-64. DOI: <https://doi.org/10.21742/ijst.2021.9.2.04>.
- [8] Girdhar Gopal Ladha, Ravi Singh Pippal. An efficient distance estimation and centroid selection based on k-means clustering for small and large dataset. *International journal of advanced technology and engineering exploration*, 2020, 7(73): 234-240. DOI: <https://doi.org/10.19101/ijatee.2020.762109>.
- [9] Xiancheng Xiahou, Yoshio Harada. B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM. *Journal of Theoretical and Applied Electronic Commerce Research*, 2022, 17(2): 458-475. DOI: <https://doi.org/10.3390/jtaer17020024>.
- [10] Feng Ye. Green Progress of Cross-border E-Commerce Industry Utilizing Random Forest Algorithm and Panel Tobit Model. *Applied Artificial Intelligence*, 2023, 37(1). DOI: <https://doi.org/10.1080/08839514.2023.2219561>.
- [11] Mengyuan Li. Research on the prediction of e-commerce platform user churn based on Random Forest model. *2022 3rd International Conference on Computer Science and Management Technology (ICCSMT)*, Shanghai, China, 2022, 34-39. DOI: <https://doi.org/10.1109/iccsmt58129.2022.00014>.

THE OPTIMAL YIELD PROBLEM OF CROP PLANTING BASED ON LINEAR PROGRAMMING MODEL

Jie Deng¹, Jian Wang¹, Hao Xu¹, ZhengBo Li^{2*}

¹Department of Brewing Engineering, MouTai Institute, Renhuai 564500, GuiZhou, China.

²Department of Public Basic Education, MouTai Institute, Renhuai 564500, GuiZhou, China.

Corresponding Author: ZhengBo Li, Email: wxxtedu@sina.cn

Abstract: The optimal income problem of crop planting is controlled by various uncertain factors. By controlling for uncertain factors, the problem of crop income can be summarized as the optimization problem of crop planting structure. Reasonable optimization of crop planting structure has an important impact on the economic status of regional farmers, sustainable development of agriculture, and sustainable utilization of land resources. The crop planting structure includes both temporal and spatial structures. Therefore, this article constructs a mathematical model combining linear programming and Monte Carlo method by analyzing the two types of land structures. Firstly, use linear programming to simulate various constraints in crop planting structure, and then increase the randomness in time structure through Monte Carlo method. At the same time, this article also explores the correlation between crop yield per mu, planting cost, and sales price through heat maps, in order to help decision-makers better analyze crop planting structure and improve crop planting income. The results indicate that optimizing the crop planting structure can improve crop yields and provide reference for decision-makers in crop planting planning and market pricing.

Keywords: Crop planting structure; Linear programming; Monte carlo method; Heat map correlation

1 INTRODUCTION

Ling Qiu County, Datong City, Shanxi Province, China has 12 townships and 186 administrative villages, with a total population of 250000, including 200000 agricultural population. Since 2013, the county has formulated the "Implementation Plan for Organic Agriculture Park in Lingqiu County (2013-2030)". After nearly 11 years of development, its total organic agriculture output value has reached 580 million yuan, and it has built the largest contiguous organic agriculture park in China. In addition, Lingqiu County has been selected as a "Shanxi Province Agricultural Product Quality and Safety Demonstration Creation Unit", and Hongshileng Township has also been rated as a "China Organic Agriculture Development Demonstration Township". The area attaches great importance to soil improvement and protection, and through research on crop planting structure and scientific and reasonable farming methods, it has laid the foundation for the sustainability of organic agriculture.

Therefore, reasonable planning of crop planting structure plays an important role in regional crop income. At present, scholars at home and abroad have conducted research on crop planting structure. Hu M et al[1] optimized the crop planting structure in Heilongjiang Province through multi-objective interval parameter planning and other methods. The results showed that slope, population density, and average temperature in the coldest season were the main factors affecting the distribution of rice, corn, and soybeans. Liu Q et al[2] established a multi-objective spatial CPSO model and conducted a case study on the middle and upper reaches of the Heihe River Basin in Gansu Province, China. The optimization of planting structure significantly improved regional water resources and ecological benefits at different scales. Adamo T et al[3] solved the intercropping system in crop planting structure based on constraint programming models of integer and interval variables. Alotaibi A et al[4] analyzed the application of feed mixing, crop patterns and rotation plans, irrigation water and product conversion through linear programming models. Adeyemo J et al[5] applied differential evolution algorithm (DE) and linear programming model (LP) to optimize planting area and maximize the use of irrigation water. Li M et al[6] conducted multidimensional optimization of AWLR (Water Resources, Land Resources, and Sustainable Development) by establishing a framework model that combines multiple models such as multi-objective programming and linear programming. Abdelwahab et al[7] studied the planting patterns in the eastern delta region, especially in the areas supplied by the Ismailia Canal, by establishing a linear programming model to solve the problem of balancing limited freshwater supply. Reddy D J et al[8] used machine learning (ML) to estimate crop yield based on weather conditions. Luo N et al[9] analyzed data from 87 field experiments in China by combining data-driven prediction with machine learning methods, and concluded that by optimizing the dense planting structure of crops, China's corn yield will increase by 52% by the 2030s. Gebre et al[10] conducted a systematic literature review of 69 articles on Multi Criteria Decision Making (MCDM) and found that Linear Programming (LP) and Simulated Annealing (SA) methods are mainly used for optimizing multi-objective complex agricultural and forest land allocation problems.

Most scholars have analyzed the crop planting structure of the block area through mathematical models, and made current and future characteristics and trends of crop planting structure in the area, or analyzed a single constraint condition in the crop planting structure. However, they rarely combine multiple constraints of crop planting structure and apply them to the problem of crop planting income.

The spatial structure of crop planting mainly includes the planting area, planting yield, and planting ratio of different

crops, while the temporal structure of crops mainly includes multiple constraints such as land rotation system, climate, crop spacing, and land use mode. The linear programming model, as an efficient mathematical model, is widely used in dealing with large-scale problems and can effectively handle a series of constraints. At the same time, for some unpredictable factors in the time structure, Monte Carlo simulation can be used to simulate unpredictable factors such as climate and market fluctuations in prices.

Therefore, based on the planting situation in Xiahe Village, Lingqiu County, Datong City, Shanxi Province in 2023, this article analyzes and optimizes the planting structure of its crops, and presents the planting situation from 2023 to 2030. By constructing a maximum profit objective function and combining linear programming and Monte Carlo methods to optimize crop planting structure, a correlation analysis was conducted on the sales cost, sales price, and yield per mu of crops in Xiahe Village in 2023.

This article combines multiple constraints in crop planting structure, such as rotation system, planting density, market fluctuations, etc., and conducts comprehensive optimization through linear programming and Monte Carlo methods. This comprehensive treatment of multiple constraint conditions provides decision-makers with a more comprehensive planting planning scheme, which can effectively improve the economic benefits of crop planting. At the same time, the Monte Carlo method was used to simulate unpredictable factors such as market volatility and climate change, enhancing the robustness and practicality of the model. This method not only helps decision-makers better cope with uncertainty, but also provides new ideas for risk management in agricultural planting.

2 APPLICATION METHOD DESCRIPTION

2.1 Linear Programming Model

The linear programming model was proposed by George Dantzig[11]. It gradually improves the solution through iteration until the optimal solution is found, which is applicable to most practical problems, but due to its own limitations, it cannot solve nonlinear problems. This model mainly consists of decision variables, objective functions, and constraint conditions.

2.1.1 Decision variables

The decision variables in linear programming models are the unknowns that need to be solved, usually represented by x_1, x_2, \dots, x_n . For example, the production quantity of each production level in the production plan.

2.1.2 Objective function

The objective function in linear programming is the function that needs to be maximized or minimized, and its general form is:

$$\text{Maximize(or Minimize)} \quad Z = c_1x_1 + c_2x_2 + \dots + c_nx_n \quad (1)$$

Among them c_1, c_2, \dots, c_n are coefficients representing the contribution of each variable in the objective function.

2.1.3 Constraints

Linear inequalities or equations that limit the values of decision variables.

General form:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n &\leq b_1 \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n &\geq b_2 \\ &\dots \\ a_{m1}x_1 + a_{m2}x_2 + \dots + a_{mn}x_n &\leq b_m \end{aligned} \quad (2)$$

Among them, a_{ij} is a coefficient and b_{ij} is a constant.

2.2 Monte Carlo Method

Monte Carlo simulation is a statistical sampling method used to evaluate solutions to quantitative problems[12]. Its theoretical basis is the law of large numbers and the central limit theorem, and one of its core ideas is to randomly select points from a certain range, and then approximate the problem through the statistical properties of these random points. For example, calculating area, calculating pi, etc. The most typical example is to randomly replace a square with a side length of 2, and then calculate the proportion of points falling into the inscribed circle. Based on this proportion, the pi can be obtained.

This article sets the probabilities of some unpredictable factors and uses Monte Carlo method to select a random value for each probability to repeatedly simulate the model, thereby achieving the randomness of unpredictable factors.

3 DATA ANALYSIS AND VISUALIZATION

Xia Che He Village contains 1216 acres of arable land, consisting of 6 land types divided into 34 plots, including flat dry land, terraced fields, hillside land, and irrigated land. These lands are suitable for growing different crops such as grains, rice, and vegetables. In addition, the village also has 16 standard greenhouses and 4 smart greenhouses, suitable for growing vegetables and edible fungi. (Data source: Shanxi Statistical Yearbook) Different crops have different planting costs, yields per mu, and sales unit prices on different land types, and some crops can only be planted on specific types of land.

3.1 Comparison of Economic Benefits of Six Land Types

Table 1 Land Area

Flat dry land(mu)	Terraced Fields(mu)	Hillside land(mu)	Irrigated land(mu)	Ordinary greenhouse(mu)	Smart greenhouse(mu)
365	619	108	109	9.6	5.4

Table 1 shows the land area of six land types. Although the crop planting cost and yield per mu of the latter three land types are higher than the first three, their planting area is lower. Here, this paper can discuss the economic benefits of each type of land to intuitively reflect which type of land would yield more ideal income from planting crops. Thus providing reference for the results of linear programming models.

The box plot provides a visual representation of the distribution and median of planting costs, yield per mu, and sales unit price for different crops in the six land types. Generally, total profit = revenue – cost, where in this article, revenue = yield per mu x sowing area x sales unit price. Here, this paper use the median in the box plot as the planting cost, yield per mu, and sales unit price for each plot type to roughly calculate and compare the economic benefits of the six land types.

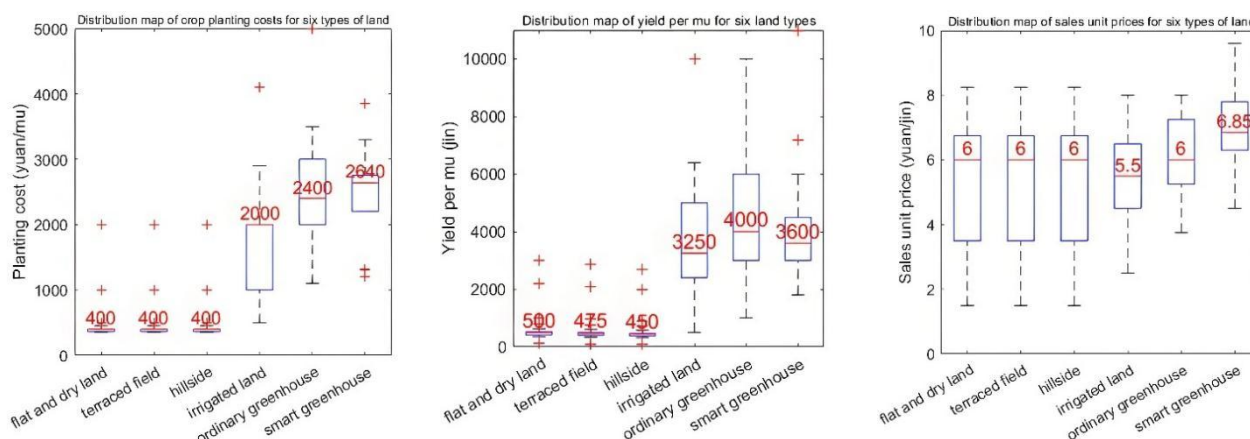


Figure 1 Distribution of Crop Planting Costs, Yield Per Mu, and Sales Unit Price among Six Land Types

According to the distribution chart of crop planting cost, yield per mu, and sales unit price among six land types in Figure 1, it suggests that the cost of smart greenhouses and ordinary greenhouses is higher in planting cost, while the planting cost of flat land, terraced fields, and mountainous areas is lower. In terms of yield per mu, irrigated land, ordinary greenhouses, and smart greenhouses have higher yields, while flat land, terraced fields, and mountainous areas have lower yields per mu. The price of smart greenhouses is higher in the sales unit price.

At the same time, this paper calculates the economic benefits of each land type by obtaining the median of planting costs, yield per mu, and sales unit price for the six land types.

Table 2 Total Profit of Six Land Types

Flat dry land(yuan)	Terraced Fields(yuan)	Hillside land(yuan)	Irrigated land(yuan)	Ordinary greenhouse(yuan)	Smart greenhouse(yuan)
949000	1516550	248400	1730375	207360	118908

According to Table 2, the total profit of the six land types shows that irrigated land has the highest economic benefits, while smart greenhouses have the lowest economic benefits. At the same time, this paper studies the correlation between the planting cost, yield per mu, and sales unit price data of crops, providing reference for decision-makers to set sales unit prices and choose crops to plant.

3.2 Correlation Analysis of Crop Planting Cost, Yield Per Mu, and Sales Unit Price

In the actual production process, total profit is related to sales price, sales cost, and yield per mu. This paper considers using Pearson correlation coefficient r to verify the internal relationship between the three. Obtain the Pearson correlation coefficient heatmap of matrices C_{ij} , P_{ij} and R_{ij} in Figure 2.

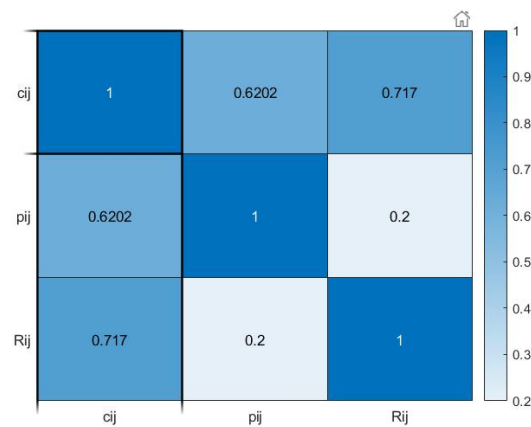


Figure 2 Pearson correlation coefficient heatmap of matrices C_{ij} , P_{ij} and R_{ij}

Finally, the correlation coefficient between C_{ij} and P_{ij} is 0.620, the correlation coefficient between C_{ij} and R_{ij} is 0.717, and the correlation coefficient between P_{ij} and R_{ij} is 0.2.

That is to say, there is a strong correlation between sales cost, sales unit price, and yield per mu. The sales unit price has a strong correlation with sales costs and a weak correlation with yield per mu. The yield per mu is strongly correlated with sales costs, but weakly correlated with sales unit prices.

When pursuing maximum profits in practice, this model needs to consider the relationship between sales unit price, sales cost, and yield per mu, rather than solely relying on the maximum or minimum value of a variable to determine how to plant a certain crop on a certain plot of land.

Based on the above data analysis, the first constraint this paper needs to determine is that different plots require different types of crops to be planted, and the planting costs, yield per mu, and sales unit price of some crops planted on different plots are also different. At the same time, it is necessary to consider the correlation between these three factors when constructing the model.

4 CONSTRUCTION OF LINEAR PROGRAMMING MODEL

4.1 Determination of Constraints on Crop Planting Structure

The optimization of crop planting structure includes multiple constraints such as variety selection, rotation system, and planting density. Based on the planting situation, soil conditions, and natural conditions in Xiahe Village in 2023, a linear programming model is constructed by considering crop rotation system, planting density, planting area of various crops, and market fluctuations as the main constraints, and establishing the objective function of maximizing income. The maximization of income under the optimization of crop planting structure is discussed.

4.1.1 Crop rotation system

According to the growth pattern of crops, each planting process absorbs trace elements from the soil, resulting in an imbalance of trace elements in the soil. Therefore, each crop cannot be continuously planted in the same plot (including greenhouses), otherwise it will reduce production. At the same time, intercropping can promote the absorption of phosphorus by crops and increase the content of phosphorus in the soil[13]. For example, the phosphorus produced by legume crops can be used as a nitrogen source in the soil, which is beneficial for the growth of other crops. Moreover, the mixed planting of legumes and grasses is highly valued in many parts of the world for its advantages of increasing yield, reducing soil erosion, and minimizing pests and diseases. So it is possible to plant legumes as much as possible during the crop planting process, which is beneficial for the economic benefits of crops. However, it should also be noted that leguminous crops cannot be planted continuously[14].

Due to the varying drainage and soil fertility indices of different food crops, implementing a rotation planting plan is necessary to avoid soil nutrient imbalance caused by long-term planting of a single crop, disrupt the growth and activity space of pests and diseases, and ensure the healthy growth of crops[15].

4.1.2 Planting area

At the same time, the planting plan should take into account the convenience of farming operations and field management. The planting areas for each crop per season should not be too scattered, and the planting area for each crop on a single plot (including greenhouses) should not be too small. Moreover, crops often need to be planted under certain temperature, humidity, and light conditions, but different crops have certain differences in the above indicators, and the sowing time is also different. Therefore, the cost of planting different crops on different land types varies, and their selling prices are also different. Therefore, it is necessary to choose a suitable planting area to achieve maximum profits.

4.1.3 Market volatility

Under market fluctuations, crop prices are influenced by numerous factors, such as seasonal fluctuations in vegetable prices caused by the production of different varieties of vegetables, inflation affecting crop price fluctuations, and the impact of external factors on agricultural product market fluctuations. For example, higher agricultural inflation may

disrupt stability through lower output and higher overall inflation[16].

Therefore, under ideal conditions where market volatility tends to be relatively stable, this model can simulate small fluctuations in market volatility by selectively increasing our variables such as expected sales volume, planting cost, yield per mu, and sales price through Monte Carlo analysis. At the same time, this paper introduces market volatility to simulate other factors that affect market volatility, such as inflation.

4.1.4 The replaceability and complementarity of crops

There may be certain substitutability and complementarity between various crops in the market. For substitutability and complementarity, this paper can introduce substitutability coefficient and complementarity coefficient to quantify the relationship between the two crops. Finally, the optimal solution can be obtained by using linear programming.

4.1.5 Expected sales volume

This paper sets the expected sales volume for each crop based on the planting plan for the 2023 vehicle and village crops, and treat it as unsold waste if it exceeds the expected sales volume.

4.2 Symbol Explanation and Model Assumptions

Table 3 Symbol Explanation

Symbol	Explanation
X_{ijt}	The planting area of crop i in season t on land j
C_{ij}	The cost of crops on j plots of land
P_{ij}	The selling price of crop i on land j
S_i	Expected sales volume of crop i
α	Minimum planting restriction coefficient
β	Penalty Coefficient
R_{ij}	Indicate the yield per mu of crop i on land j
δ_{kT}	Indicate the increase or decrease rate of the k -th option in the T -th year
Q_{ijt}	Indicate the yield of crop i on land j in the t -th quarter
B_{jt}	Indicate the area of j land in quarter t

According to Table 3, it is easy to understand the meaning of the formulas in the following text. For the impact of market fluctuations and other factors on crops, this chapter plans to optimize the crop planting structure for 7 years, and adjust the range of expected sales volume, planting costs, yield per mu, and sales price fluctuations annually.

Expected sales volume: The annual growth rate of wheat and corn ranges from 5% to 10%, while the expected sales volume of other crops fluctuates by $\pm 5\%$ relative to 2023.

Planting cost: All crops increase by approximately $\pm 5\%$ annually.

Mu yield: The annual mu yield of all crops may fluctuate by $\pm 10\%$.

Sales price: The prices of grain crops are basically stable. For vegetable crops, the sales unit price increases by about 5% annually, while the sales unit price of edible mushrooms decreases by 1% to 5% annually, especially the sales unit price of morel mushrooms decreases by 5% annually.

4.3 Establishment of Objective Function and Constraint Conditions

4.3.1 Objective function

$$\sum_{t=1}^{14} \sum_{j=1}^{54} \sum_{i=1}^{41} P_{ij} \times R_{ij} \times X_{ijt} - C_{ij} \times X_{ijt} \quad (3)$$

Among them, i , j , and t are all integers (the same applies in the following text)

4.3.2 Constraints

1. The planting area of each crop cannot exceed the area on that land.

$$\sum_{t=1}^{14} \sum_{j=1}^{54} \sum_{i=1}^{41} X_{ijt} \leq B_{jt} \quad (4)$$

Among them, B_{jt} represents the area of the j -th piece of land in the t -th quarter

2. Grain crops can only be planted in flat dry land, terraced fields, and mountain slopes, except for rice. Vegetable crops can be planted in irrigated land and two types of greenhouses, while edible fungi can only be planted in ordinary greenhouses.

In flat land, terraced fields, and hillside areas, there are:

$$\sum_{i=16}^{41} X_{ijt} = 0, j \in [1,26], t \in [1,14] \quad (5)$$

In irrigated land, there are:

In the first quarter of irrigated land, only crops with $i \in [17,34]$ can be planted, and in the second quarter, only crops with $i \in [35,37]$ can be planted.

$$\sum_{i=1}^{16} X_{ijt} + \sum_{i=38}^{41} X_{ijt} = 0, j \in [27,34], t \in [1,14] \quad (6)$$

$$\sum_{i=35}^{37} X_{ijt} = 0, j \in [27,34], t \in \{1,3, \dots, 11\} \quad (7)$$

$$\sum_{i=17}^{34} X_{ijt} = 0, j \in [27,34], t \in \{2,4, \dots, 14\} \quad (8)$$

In ordinary greenhouses, there are:

In ordinary greenhouses, only crops from 17 to 34 can be selected for planting in the first season, and only crops from 38 to 41 can be selected for planting in the second season.

$$\sum_{i=1}^{16} X_{ijt} + \sum_{i=35}^{37} X_{ijt} = 0, j \in [35,50], t \in [1,14] \quad (9)$$

$$\sum_{i=38}^{41} X_{ijt} = 0, j \in [27,34], t \in \{1,3, \dots, 11\} \quad (10)$$

$$\sum_{i=17}^{34} X_{ijt} = 0, j \in [27,34], t \in \{2,4, \dots, 14\} \quad (11)$$

In the smart greenhouse, there are:

The smart greenhouse can plant two seasons of vegetables every year (excluding Chinese cabbage, white radish, and red radish). Moreover, Chinese cabbage, white radish, and red radish can only be planted in the second season.

$$\sum_{i=1}^{16} X_{ijt} + \sum_{i=38}^{41} X_{ijt} = 0, j \in [51,54], t \in [1,14] \quad (12)$$

$$\sum_{i=35}^{37} X_{ijt} = 0, j \in [51,54], t \in \{1,3, \dots, 11\} \quad (13)$$

3. Leguminous crops must be planted every three years.

$$\sum_{t=1}^6 \sum_{i=1}^5 X_{ijt} + \sum_{t=1}^6 \sum_{i=17}^{19} X_{ijt} \geq 1, \forall j \quad (14)$$

$$\sum_{t=6}^{12} \sum_{i=1}^5 X_{ijt} + \sum_{t=6}^{12} \sum_{i=17}^{19} X_{ijt} \geq 1, \forall j \quad (15)$$

4. Crop management requires that the planting area of crops should not be too small (Planting area).

$$X_{ijt} \geq \alpha, \forall i, j, t \quad (16)$$

Among them, α is the minimum limit constant for planting area.

5. Market fluctuations

We use Monte Carlo method to simulate the changes in market volatility within the following range of variation. There is a growth trend of 5% to 10% for wheat and corn. Other crops have a growth trend of $\pm 5\%$. The sales unit price of grain crops is basically stable, while the sales unit price of vegetable crops has a growth trend of 5%. Edible mushrooms decrease by about 1% to 5% annually, with more mushrooms decreasing by 5%. The yield per mu of all crops varies by $\pm 10\%$ annually. The planting cost of all crops increases by 5% annually.

Changes in expected sales volume, unit price, yield per mu, and planting cost on the original data. It can be divided into the following 7 schemes.

(1) The expected sales volume of wheat and corn is:

$$S_i \times (1 + \delta) = S_i', i = 6,7, \delta \in [0.05, 0.10] \quad (17)$$

(2) The expected sales volume of other crops is:

$$S_i \times (1 + \delta) = S_i', i \neq 6,7, \delta \in [-0.05, 0.05] \quad (18)$$

(3) Sales unit price of grain crops:

$$P_{ij} = P_{ij}', i = [1,16] \quad (19)$$

(4) Sales unit price of vegetable crops:

$$P_{ij} \times (1 + \delta) = P_{ij}', i = [17,37], \delta \in [0, 0.05] \quad (20)$$

(5) Sales unit price of edible mushrooms:

$$P_{ij} \times (1 + \delta) = P_{ij}', i = [38,40], \delta \in [-0.05, -0.01] \quad (21)$$

$$P_{ij} \times (1 + \delta) = P_{ij}', i = 41, \delta = -0.05 \quad (22)$$

(6) The yield per mu of all crops:

$$R_{ij} \times (1 + \delta) = R_{ij}', \forall i, \delta \in [-0.1, 0.1] \quad (23)$$

(7) The planting cost of all crops:

$$C_{ij} \times (1 + \delta) = C'_{ij}, \forall i, \delta \in [0, 0.05] \quad (24)$$

For cases (1)~(7), take random numbers of δ within the range of δ for each case, because there are a total of 7 years, so each case needs to take 7 random numbers to achieve the impact of unpredictable factors on the planting plan. The value of δ for the k th scheme in the T th year is denoted as δ_{kT} .

Among them, $k = [1, 7]$, $T = [1, 7]$.

6. The complementarity and substitutability of crops

Complementarity explains the mutual influence between different crops, and this paper set a complementarity coefficient to measure the complementarity between two different crops.

$$Q_{ijt} = X_{ijt} \times R_{ij} \times (1 + \sum_{k=1}^{41} \omega_{ik} \times Q_{kjt}) \quad (25)$$

Among them: ω_{ik} , Indicating the complementary strength of the i th crop to the k th crop, that is, planting the i th crop will increase the yield of the k th crop. Q_{ijt} represents the yield of the i -th crop in the j th quarter of the j th plot of land. Yield = planting area \times yield per mu.

The substitutability indicates that the two crops have an opposite relationship, that is, an increase in crop i 's yield will result in a decrease in crop k . This paper uses substitution coefficients to measure this relationship.

$$Q_{ijt} = X_{ijt} \times R_{ij} \times (1 + \sum_{k=1}^{41} \theta_{ik} \times Q_{kjt}) \quad (26)$$

Among them, θ_{ik} represents the substitution coefficient between the i th crop and the k th crop, and the stronger the coefficient, the more the k th crop reduces the yield of i .

7. Market price fluctuations

Our net income is determined by market volatility and selling price, when the market volatility is τ . So our earnings are:

$$P_{ij} \times \min(Q_{ijt}, S_i) \times (1 + \tau) \quad (27)$$

Among them, $\min(Q_{ijt}, S_i)$ represents our handling of unsold and wasteful excess parts, and τ is the market volatility.

4.4 Experimental Results and Discussion

Through the correlation analysis of crop planting structure, yield per mu, and sales unit price, this paper finally has obtained the Pearson correlation coefficient of the three as shown in the Table 4 below.

Table 4 Correlation coefficients between planting cost, yield per mu, and sales unit price.

C_{ij} and R_{ij}	C_{ij} and P_{ij}	R_{ij} and P_{ij}
0.717	0.620	0.200

By constructing a linear programming model and Monte Carlo method to simulate unpredictable factors, and analyzing the constraints of crop planting structure on crop planting area, rotation system, crop selection, market fluctuations, weather prediction, and legume rotation, as well as constructing objective functions, a linear programming model was obtained, resulting in a total profit of 270 million yuan after 7 years of simulation.

The use of linear programming models and Monte Carlo methods to solve such problems provides methods and references for decision-makers in actual production processes. Simultaneously using Monte Carlo method to quantitatively analyze some unpredictable factors can effectively avoid the risks that may lead to reduced returns in the actual production process, as well as other related risks. In the correlation analysis, the correlation between crop planting cost, yield per mu, and sales unit price was also explained, among which planting cost and yield per mu, planting cost and sales unit price have a strong correlation.

5 CONCLUSION AND IMPLICATIONS

With the increasing demand for sustainable agricultural development, reasonable crop planting strategies are crucial for improving production efficiency, reducing planting risks, and achieving efficient resource utilization.

Starting from the issue of crop planting structure, this article conducted data analysis and visualization on 54 plots of 41 crops and 6 types of land. Research has found that different crops require different restrictions to be planted on different lands. At the same time, constraints such as planting legumes and crop rotation in actual production were added to the model to make it more realistic. Then, through model assumptions, the range of market volatility was set. Finally, construct the objective function and constraints of linear programming to obtain the final yield of 7-year planting.

The crop planting structure optimization model based on linear programming and Monte Carlo method proposed in this article has strong application feasibility and wide applicability. This model can effectively help decision-makers optimize the planting structure and improve the economic benefits of crop planting by comprehensively considering multiple constraints such as planting costs, yield per mu, and sales prices. Especially in rural areas of China, this model can provide farmers with scientific planting planning, reduce planting risks, and improve land use efficiency. In addition, the model also simulated unpredictable factors such as market volatility and climate change through Monte Carlo

method, enhancing the robustness and practicality of the model.

Future research directions can further expand the applicability of the model, such as introducing nonlinear programming methods to address more complex agricultural problems, or combining big data and artificial intelligence technologies to adjust planting strategies in real time. In addition, the model can consider more environmental factors such as natural disasters, pests and diseases to improve the accuracy of predictions. Finally, the model can also be applied to other agricultural fields such as animal husbandry, fisheries, etc., further promoting the sustainable development of agriculture.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

FUNDING

The research is funded by the projects: Research on the Characteristic Education of Advanced Mathematics from the Perspective of HPM(GZLCLH-2020-172), Exploration and Practice of Ideological and Political Teaching Reform in the Advanced Mathematics Course(2023383), Mathematics Innovation and Integration Teaching and Research Group, and Course Ideological and Political Demonstration Course "Linear Algebra".

REFERENCES

- [1] Hu M, Tang H, Yu Q, et al. A new approach for spatial optimization of crop planting structure to balance economic and environmental benefits. *Sustainable Production and Consumption*, 2025, 53: 109-124.
- [2] Liu Q, Niu J, Du T, et al. A full-scale optimization of a crop spatial planting structure and its associated effects. *Engineering*, 2023, 28: 139-152.
- [3] Adamo T, Colizzi L, Dimauro G, et al. Crop planting layout optimization in sustainable agriculture: A constraint programming approach. *Computers and Electronics in Agriculture*, 2024, 224: 109162.
- [4] Alotaibi A, Nadeem F. A review of applications of linear programming to optimize agricultural solutions. *International Journal of Information Engineering and Electronic Business*, 2021, 15(2): 11.
- [5] Adeyemo J, Otieno F. Optimizing planting areas using differential evolution (DE) and linear programming (LP). *International Journal of Physical Sciences*, 2009, 4(4): 212-220.
- [6] Li M, Cao X, Liu D, et al. Sustainable management of agricultural water and land resources under changing climate and socio-economic conditions: A multi-dimensional optimization approach. *Agricultural Water Management*, 2022, 259: 107235.
- [7] Abdelwahab H F, Negm A M, Ramadan E M, et al. Mitigating water shortages and enhancing food security through crop optimization: Insights from the Eastern Nile Delta. 2024.
- [8] Reddy D J, Kumar M R. Crop yield prediction using machine learning algorithm//2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS). IEEE, 2021: 1466-1470.
- [9] Luo N, Meng Q, Feng P, et al. China can be self-sufficient in maize production by 2030 with optimal crop management. *Nature Communications*, 2023, 14(1): 2637.
- [10] Gebre S L, Cattrysse D, Alemayehu E, et al. Multicriteria decision making methods to address rural land allocation problems: A systematic review. *International Soil and Water Conservation Research*, 2021, 9(4): 490-501.
- [11] Dantzig G B. Linear programming. *Operations research*, 2002, 50(1): 42-47.
- [12] Pramesti F A, Subekti H, Candra A D. Application of Monte Carlo simulation for the estimation of production availability in geothermal well//IOP Conference Series: Earth and Environmental Science. IOP Publishing, 2024, 1339(1): 012017.
- [13] Zhu S G, Cheng Z G, Wang J, et al. Soil phosphorus availability and utilization are mediated by plant facilitation via rhizosphere interactions in an intercropping system. *European Journal of Agronomy*, 2023, 142: 126679.
- [14] Ma L, Ma S, Chen G, et al. Mechanisms and mitigation strategies for the occurrence of continuous cropping obstacles of legumes in China. *Agronomy*, 2023, 14(1): 104.
- [15] Moldavan L, Pimenowa O, Wasilewski M, et al. Crop rotation management in the context of sustainable development of agriculture in Ukraine. *Agriculture*, 2024, 14(6): 934.
- [16] Ginn W. Agricultural fluctuations and global economic conditions. *Review of World Economics*, 2024, 160(3): 1037-1056.

THE CROP PLANTING PROBLEM BASED ON THE OPTIMIZATION MODEL

YiTong Liu^{1*}, YiLin Wang², JiLing Zou³

¹Department of business, Accounting, Xi'an International Studies University, Shaanxi, 710128, Xi'an, China.

²Department of business, Business English, Xi'an International Studies University, Shaanxi, 710128, Xi'an, China.

³Department of economics and finance, finance, Xi'an International Studies University, Shaanxi, 710128, Xi'an, China.

Corresponding Author: YiTong Liu, Email: 18991002666@163.com

Abstract: The development of modern agricultural technology and the limited cultivated land planting resources in China make people pay more attention to the choice of planting plan. In the process of crop planting, whether farmers can comprehensively consider the impact of market price fluctuations, climate change and related commodities (substitutes and complementary products) on the sales unit price, and then make reasonable prediction and estimation of sales volume and output of agricultural products, which has far-reaching significance for the correct choice of planting plan. At the same time on the basis of fully considering the unique properties of different terrain, need to study the diversity of crop species and the adaptation of environment, clever use of refinement means such as rotation and terrain partition, constantly optimize planting layout strategy, and implement dynamic adjustment mechanism, to ensure that planting scheme can follow the environmental change, accurately realize the maximum maximization under the established conditions. In order to meet the farmers' yearning for a better life, this paper reasonably arranges the replacement planting of the optimal crops on various types of land in different seasons, which can effectively balance the unsalable sales and reduced sales caused by the large gap between the expected sales volume and the actual production volume, so as to maximize the annual net income. In view of the uncertainty of crop market and yield, especially the dynamic balance between per-mu yield and expected sales volume, this paper proposes targeted planting adjustment strategies to ensure the effectiveness of planting scheme and the stability of income.

Keywords: Optimization model; Crop planting; Sustainable development; Regression model

1 INTRODUCTION

In the research field of crop planting models, numerous previous studies have primarily focused on certain key factors, such as soil fertility and basic market demands [1,2]. These works have laid a foundation for agricultural planting planning. However, with the increasing complexity of the agricultural environment [3], their limitations have become more prominent. They often fail to comprehensively consider market price fluctuations, climate change, and the intricate relationships among crops, including substitution and complementarity. As a result, the planting recommendations provided lack the necessary specificity and adaptability to meet the diverse needs of farmers in different planting scenarios.

This paper aims to address these limitations and optimize crop planting strategies. By integrating agricultural knowledge with mathematical modeling and data - driven methods, a comprehensive analysis of multiple factors is conducted. The research focuses on the period from 2024 - 2030, aiming to optimize crop planting schemes while considering the dynamic balance between per - mu yield and expected sales volume. The growth rate or decline rate involved in the existing literature is given in the form of interval, and the exact value is not found for accurate calculation and approximation error in analysis, and key details such as the specific value of the planting area and the degree of spatial dispersion are not fully considered. To further verify the superiority and applicability of the established model, this paper compares it with the traditional optimization model of crop planting scheme. Traditional models often consider only a single factor or a few factors, while the model established in this paper considers many factors, including market price fluctuations, climate change, substitution and complementarity among crops. Through comparison, the model established in this paper in the optimization of planting scheme, improve economic benefit has obvious advantages, specifically, the traditional model often can only give relatively general planting advice, and the model established in this paper can according to the specific situation of different plots, different crops, give more detailed and specific planting scheme, so as to better meet the actual needs of farmers.

This paper defines key assessment dimensions considering different land types, using an identification system from A to F, and takes farmers' actual interests into account. A model is established to calculate the annual profit of crop sales, considering variables like annual crop sales volume, per mu yield, and planting costs. Different situations based on the relationship between per - mu yield and expected sales volume are analyzed, with risk coefficients introduced to account for uncertainties. The influence of commodity correlations on the planting scheme is explored through a linear programming model, considering substitution and complementarity coefficients. Through these efforts, this paper aims to help farmers make more scientific planting decisions, maximize their annual net income, and contribute to the sustainable development of agriculture [4]. On the basis of the basic planning scheme, this paper further considers and improves the fluctuation of the risk coefficient, enhances the application of the scheme, and objectively considers the

impact of the relationship between the per-mu yield and the expected sales volume on the profit amount, which is in line with the actual situation. The model is based on data operation, with concise design and strong practicability.

2 THE OPTIMIZATION OF 2024 ~ 2030 CROP PLANTING SCHEME

2.1 Variable Declaration

Integrating agricultural knowledge, in this paper, exploring the optimal path to optimize planting strategy, it is supposed to comprehensively balance the constraints of cultivated land area with the specific environmental requirements of crop growth [5,6]. Based on these considerations, this paper initially defines the following key assessment dimensions. For land type factors, the identification system from A to F can be adopted to correspond to the plots in dry land, terraces, hillsides, irrigated land, ordinary greenhouses, and smart greenhouses, deeply concerned about the actual interests and expectations of farmers. In this paper, the annual profit S_m from 2024 to 2030 is set as the core standard to measure the success of the planting scheme. In order to accurately grasp the multiple factors affecting the annual profit S_m , this paper further analyzes the key variables such as annual crop sales volume, q_j , F_j , per mu yield, Q_m , annual total planting cost C_m , crop unit planting cost C_j , planting area α_j and crop unit selling price y_j .

2.2 Optimization Of The Planting Scheme

2.2.1 Model establishment

After considering multiple influencing factors [7], The formula for calculating the profit of crop sales can be derived. The core of this formula is the difference between profit and cost. Further establish the following formula:

$$\begin{cases} \sum_{i=1}^n S_m = \sum_{i=1}^n Q_m - \sum_{i=1}^n C_m \\ Q_m = q_j \times F_j, C_m = C_j \times \alpha_j \end{cases} \quad (1)$$

Similarly, when the excess portion is sold at 50% of the 2023 sales price, the formula is as follows:

$$\sum_{i=1}^n S_m = \sum_{i=1}^n 0.5q_j(F_j + y_j) \quad (2)$$

2.2.2 Solutions

In order to prevent production reduction, each crop cannot be continuously repeated on the same plot, and to ensure soil conditions, all types of plots must be grown every three years. According to the above formula, the following conclusions are drawn as shown in Table 1:

Table 1 Six Optimal Crops of Ground Types(1)

Place	Season 1 Optimal Crop	Season 2 Optimal Crop	Optimal Bean Crop
Flat Dry Land			
Bench Terrace		No.1 Sweet Potato	
Shoulder		No.2 Buckwheat	Black Soya Bean
Irrigable Land	Cucumber	Cabbage	
Ordinary Greenhouses	Cucumber	Mushroom	Cowpea
Smart Greenhouse		No.1 Cucumber No.2 Water Cabbage	

In dry lands, terraces and mountain slopes, sweet potatoes are the most profitable non-legume crop, while black beans have the greatest potential among legumes. Chinese cucumber had the best yield in the second quarter; cowpeas had the best yield in the bean crop. In greenhouse planting, cowpea always leads the lead in bean crops, not affected by season and greenhouse type. The first season of the cucumber, and the second quarter of the mushroom is the best income; in the wisdom greenhouse, the best economic benefits. Similarly, the conclusion of the optimal planting scheme for bean crops is shown in Table 2:

Table 2 Six Optimal Crops of Ground Types(2)

Place	Season 1 Optimal Crop	Season 2 Optimal Crop	Optimal Bean Crop
Flat Dry Land		No.1 Buckwheat	Black Soya Bean

Terrace	No.2 Sweet Potato		
Shoulder			
Irrigable Land	Cabbage	Eggplant	
Ordinary Greenhouses	Oil Barley	Mushroom	Cowpea
Smart Greenhouse	No.1 Elm Yellow Mushroom		
	No.2 Yellow Heart Dish		

By comparison, the annual net income of the highest legume income is less than the top two non-legume crops. But to cultivate soil nutrition, legume crops will be grown at least once every three years from 2023. Therefore, minimizing bean planting in 2024-2030 is a necessary decision to improve the benefits, and this classification hypothesis is analyzed in this paper. The umes planted in 2023 should be replanted in 2026 and 2029; those not planted in 2023 should be planted in 2025 and 2028. Combining the above, the optimal planting scheme from 2024 to 2030 is finally obtained, as shown in Table 3:

Table 3 Optimal Planting Scheme from 2024 to 2030

2023	2024	2025	2026	2027	2028	2029	2030
√	No.1	No.2	√	No.1	No.2	√	No.1
-	No.1	√	No.1	No.2	√	No.1	No.2

Note: √ Indicates the year when the bean crop was grown

3 OPTIMIZATION OF PLANTING PLANS BASED ON DYNAMIC BALANCE BETWEEN YIELD PER MU AND EXPECTED SALES

When analyzing the crop planting scheme, the dynamic balance between the per-mu yield and the expected sales volume should be accurately considered, based on the calculation of the smaller values [8]. If the yield per mu volume is lower than the expected sales volume, the actual available volume is limited by the yield per mu, and the inventory cannot meet the expectation; otherwise, the part exceeding the expected sales volume will be unsalable. This paper adopts the conditional classification discussion strategy, and deduce the annual profit calculation model when the yield per mu yield is greater than or less than the expected sales volume, so as to determine and optimize the planting strategy and maximize the revenue [9].

3.1 The Per-Mu Yield $H_{xj} < \text{The Expected Sales Volume } Q_{xj}$

3.1.1 Model establishment

The profit S_{xj} is calculated as follows, in which the annual sales Q_{xj} is calculated from the per-mu yield H_{xj} .

$$\begin{cases} S_{xj} = \sum_{i=1}^n Q_{xj} - C_x, \\ C_x = \alpha_j \times \sum_{i=1}^n C_{xj}, \\ Q_{xj} = P_{xj} \times H_{xj} \times \alpha_j \end{cases} \quad (3)$$

3.1.2 Risk adjustment of mu yield

Considering the fluctuation of per-mu yield within the range of $\pm 10\%$ and its uncertainty, under the premise of per-mu yield as the key indicator affecting the annual sales, Introduce risk coefficient θ , Reexpressed the possibility of a surge or drop in the yield of a certain crop due to climate change, insect disasters and other factors. List the adjusted formula:

$$Q_{xj} = P_{xj} \times H_{xj} \times \alpha_j \times \theta \quad (4)$$

3.1.3 Fluctuation of grain crop sales price

For food crops, the annual sale price P_{xj} basically stable, so there's no need to do too much calculation.

3.1.4 Fluctuation of sales prices of vegetable crops

$$P_{xv} = P_{2023v} (1+5\%)^{x-2023} \quad (5)$$

3.1.5 The fluctuation of the sales price of edible fungus crops

Among all measurable fungus crops, the selling price of morels showed a significant decline of 5% per year, which led the fluctuation of 1% to 5% in fungus crops. Based on careful consideration of economic interests, this paper decided not to include the cultivation plan in the planning period from 2024 to 2030. Therefore, in this round of analysis, this paper will focus on evaluating and calculating the price fluctuations of other bacterial crops besides mochella, to ensure

the optimization of planting strategies and maximize profits. To this regard, the fluctuation coefficient is introduced λ , represents the extent of the price change of edible fungi.

$$P_{xm} = P_{2023m} (1 - \lambda)^{x-2023}, (0.01 < \lambda < 0.05) \quad (6)$$

3.1.6 Objective function after synthesis

$$S_{xj} = \sum_{i=1}^n \theta [P_{xc} \times H_{xc} \times \alpha_c + P_{2023v} (1 + 5\%)^{x-2023} \times H_{xv} \times \alpha_{xv} + P_{2023m} (1 - \lambda)^{x-2023} \times H_{xm} \times \alpha_m] - \sum_{i=1}^n \alpha_j \times C_{xj} \quad (7)$$

3.2 The Per-Mu Yield H_{xj} > The Expected Sales Volume Q_{xj}

$$\begin{cases} S_{xj} = \sum_{i=1}^n Q_{xj} - C_x, \\ C_x = \alpha_j \times \sum_{i=1}^n C_{xj}, \\ Q_{xj} = P_{xj} \times q_{xj} \times \alpha_j \times \theta \end{cases} \quad (8)$$

Use the expected sales q_{xj} to calculate sales Q_{xj} . At this time, the risk factors of per-mu yield and the sales price of each crop are the same as 3.1.

3.2.1 Expected sales of wheat and corn fluctuate

$$q_{xc'} = q_{2023} \times (1 \pm \mu)^{1-2023}, (0.05 < \mu < 0.1) \quad (9)$$

3.2.2 Volatility in the expected sales volume of other crops

$$q_{xo} = q_{2023} \times (1 \pm 5\%)^{1-2023} \quad (10)$$

3.2.3 Objective function after synthesis

$$S_{xj} = \sum_{i=1}^n \theta \{ P_{xc'} \times \alpha_{c'} \times [q_{2023} \times (1 \pm \mu)^{1-2023} + q_{oc}] + P_{2023v'} (1 + 5\%)^{x-2023} \times q_{xv'} \times \alpha_{v'} + P_{2023m'} (1 - \lambda)^{x-2023} \times q_{xm'} \times \alpha_{m'} \} - \sum_{i=1}^n \alpha_j \times C_{xj} \quad (11)$$

$$(0.01 < \lambda < 0.05, 0.05 < \mu < 0.1)$$

3.3 The Optimized Result

By substituting actual data into the model for solving, the optimal planting plan from 2024 - 2030 is obtained. Among grain crops, buckwheat and sweet potato each account for 30% of the planting proportion, and the remaining grain crops together account for 40%. For legume crops, only black soybeans or cowpeas are selected. Among vegetable crops, eggplant accounts for 36% and cabbage accounts for 26%, and the remaining vegetables together account for 48%. Compared with the previous plan, the planting areas of buckwheat and sweet potato increase by 52 mu and 46 mu respectively, with an increase - decrease relationship among other grain crops. The planting areas of kidney beans and cowpeas decrease by 49 mu and 26 mu respectively, while those of potatoes and cucumbers increase by 27 mu and 46 mu respectively, and the areas of other vegetables decrease slightly.

4 THE OPTIMAL PLANTING SCHEME RELATED TO COMMODITY CORRELATION

4.1 Establishment Of The Linear Programming Model

$$\sum_{i=1}^n S_m = \sum_{i=1}^n Q_m - \sum_{i=1}^n C_m \quad (12)$$

4.2 Mechanism of the Influence of the Correlation

4.2.1 Alternative influence mechanism

To measure the degree of substitution between goods, the substitution coefficient is introduced $k_{y_1y_2}$. The larger they are, the stronger the substitution. With $y_{2,x}$ represents the expected sales volume at this time, using $y_{2,x'}$ Indicates the original expected sales volume, $\alpha_{i,a,x}$ indicates the crop area planted on block I in year x a:

$$y_{2,x} = y_{2,x'} - k_{y_1y_2} \times \alpha_{i,a,x} \quad (13)$$

4.2.2 Complementary sex influence mechanism

Introducing the complementarity coefficient $\phi_{y_1y_2}$ to measure the complementarity of agricultural products, $\phi_{y_1y_2}$. The greater the complementarity, the stronger the vice versa. At this time with $Q_{2,x}$ represents the expected sales volume at this time, using $Q_{2,x'}$ indicates the original expected sales volume:

$$Q_{2,x} = Q_{2,x'} + \phi_{y_1y_2} \times \alpha_{i,a,x} \quad (14)$$

4.2.3 Correlation between sales volume and price

The increase of demand is positively correlated with the increase of price and demand, which leads a positive correlation coefficient $\omega_{p,Q}$.

$$Q_{y,x} = Q_{y,x'} + \omega_{y_1y_2} (y_{2,x} + \overline{y_{2,x}}) \quad (15)$$

4.2.4 Correlation between planting cost and price

The correlation coefficient is introduced $\nu_{c,p}$. Planting costs increase, and market prices rise. $C_{2,x}$ means the cost of planting, $\overline{C_{2,x}}$ represents the average planting cost.

$$Q_{y,x} = Q_{y,x'} + \nu_{y_1y_2} (C_{2,x} + \overline{C_{2,x}}) \quad (16)$$

4.3 The Result of The Correlation Coefficient

After data simulation, the solved correlation coefficient value is shown in the following:

$$k_{y_1y_2} = 0.25, \phi_{y_1y_2} = 0.3, \omega_{p,Q} = 0.15, \nu_{c,p} = 0.2$$

After considering the substitutability and complementarity of various crops and the correlation between expected sales volume, sales price and the cost of planting, the optimal planting strategy from 2024 to 2030 was: buckwheat and sweet potato accounted for 30% for food crops and 40%; bean crops selected black beans or cowpea; among vegetable crops, eggplant accounted for 36 plants, and other vegetables accounted for 48%.

5 CONCLUSION

This paper considers many factors such as market price fluctuation, climate change, substitution and complementarity among crops, and studies crop cultivation deeply based on the optimization model. By constructing the specific yield calculation formula and combining the parameters of risk coefficient and fluctuation coefficient, the planting scheme of different plots and different crops is optimized. The results show that non-beans such as sweet potato and cucumbers should be planted in irrigated land, which can effectively improve economic benefits. This paper considering the dynamic balance between mu yield and expected sales, and introduce the risk coefficient to reflect the influence of climate change, diseases and insect pests and the correlation between commodities, further introduce the fluctuation coefficient, the planting scheme more detailed optimization, the results show that in food crops, buckwheat and sweet potato planting proportion appropriate increase, bean crops should choose black beans or cowpea to improve the efficiency. This paper innovatively considers many practical factors, and introduces the risk coefficient and fluctuation coefficient to enhance the practicability of the model, while taking into account the influence of the correlation between commodities on the planting scheme. The research results of this paper have important guiding significance for the future crop planting, and can help farmers to make planting plans more scientifically and improve the yield and quality of crops. It is of great significance to improve the economic benefits and realize the sustainable development of agriculture, and is expected to provide more comprehensive and precise support for the agricultural development in the

future. Based on the research of this paper, the model parameters can be refined in the future to obtain more accurate planting schemes.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCE

- [1] Zhang Xuhui, Wu Haimei, Yang Rongyan, et al. Development status and countermeasures of corn seed industry in Gansu province. *China Seed Industry*, 2025(02): 38-41.
- [2] Yu Hongyang. Study on production Structure Optimization of Agricultural Planting Industry in Jilin Province based on linear planning model. *Agricultural Information of China*, 2013 (9): 290-291.
- [3] Cheng Qiuying. Price fluctuation of agricultural products, market linkage and consumption welfare effect of rural residents. *Business Economics Research*, 2024, (15): 101-104.
- [4] Yuan Y ,Wang J ,Gao X , et al. Optimizing planting management practices considering a suite of crop water footprint indicators — A case-study of the Fengjiashan Irrigation District. *Agricultural Water Management*, 2025: 307109261-109261.
- [5] Hu M, Tang H, Yu Q, et al. A new approach for spatial optimization of crop planting structure to balance economic and environmental benefits. *Sustainable Production and Consumption*, 2025, 53109-124.
- [6] Huang Y, Guo B, Huang Y. A Study of Crop Planting Schemes Based on Linear Programming and Monte Carlo Simulation. *Journal of Innovation and Development*, 2024, 9(2): 91-96.
- [7] Zhao Rongyang, Deng Yun, Ou Gaofei. Design of big data platform for agricultural product price prediction based on index smoothing. *Internet of Things Technology*, 2024, 14(08): 113-116. DOI:10.16667/j.issn.2095-1302.2024.08.029.
- [8] Liu Z, Tan X, Li Y. Optimization of Crop Planting Strategy Based on Interior Point Method. *Agricultural & Forestry Economics and Management*, 2024, 7(2).
- [9] Zhai B, Zhu H, Wan H. Research on Crop Planting based on Linear Programming and Multiple Regression Functions. *Frontiers in Computing and Intelligent Systems*, 2024, 10(2): 43-49.

MATCHING CONTROL STRATEGY FOR HYDROGEN CIRCULATION SYSTEM FOR ON-BOARD FUEL CELLS

JianHua Liu^{1*}, Jun Li¹, JingGuang Xie¹, NanNan Liao¹, YaNan Gao²

¹CRRC QISHUYAN CO., LTD, Changzhou, 213000, Jiangsu, China.

²Dept. of Economic Research, CRRC Academy Co., Ltd, Beijing 100070, China.

Corresponding Author: JianHua Liu, Email: liujianhua.qj@crrecg.cc

Abstract: A fuel cell hydrogen recirculation system improves hydrogen utilization by circulating hydrogen from the fuel cell anode outlet to the inlet. First, by analyzing the flow rate and pressure demand of fuel cell hydrogen circulation, the model of hydrogen circulation pump is determined and the demand allocation between the inducer and the hydrogen circulation pump is carried out. Based on the one-dimensional design theory of the ejector, a one-dimensional model of the ejector is established, and by fitting the flow-pressure rise curves of the ejector and the hydrogen circulating pump, the pressures, flow rates, and cycle ratios are calculated under the two arrangement schemes of series and parallel connection, and the corresponding matching strategies are formulated. By comparing the performance of the two schemes in series and parallel with the corresponding matching strategy, the arrangement scheme of the ejector and hydrogen circulation pump is determined.

Keywords: Fuel cells; Hydrogen recirculation system; Pilot injector; Hydrogen recirculation pumps

1 INTRODUCTION

Proton Exchange Membrane Fuel Cell (PEMFC), as a key carrier for hydrogen energy utilization, provides a reaction site for hydrogen and oxygen to convert chemical energy into electrical energy, which has the advantages of clean and efficient, stable operation, fast start-up, and faster response to load demand [1], and it is the most likely replacement of internal combustion engine in future. One of the energy devices [2].

Since the PEMFC anode outlet contains hydrogen that is not involved in the reaction, if it is directly discharged, it will cause environmental pollution and even safety hazards, as well as wasting energy and increasing the cost of use [3]. In recent years, the method of recycling hydrogen from the anode outlet to the inlet for secondary hydrogen utilization using a recycling device has been widely studied and tested in order to improve fuel economy. However, the traditional single-component hydrogen recycling scheme is not mature enough for fuel cells, which mainly includes the following two aspects: first, the single pilot scheme, which has good hydrogen recycling capability in the operating zone where the output power of the electric pile is higher than 50% of the maximum power, but has poor pilot performance in the operating zone below 50% of the maximum power [4]; second, the single hydrogen recycling pump scheme, which can adjust the rotational speed according to the power of the electric pile to control the amount of hydrogen circulation, which can satisfy the circulation demand under full operating conditions, but at the same time, it will generate large parasitic power.

There are usually three schemes for fuel cell anode tail gas treatment: circulation mode, dead-end mode, and recirculation mode [5]. In the circulation mode, the gas from the anode outlet is directly discharged into the atmosphere; this scheme has a simple structure and low system cost, but the hydrogen is not fully utilized and a humidifier is required to humidify the inlet hydrogen to prevent membrane drying. In the dead-end mode, the exhaust valve is normally closed to prolong the residence time of hydrogen in the stack and thus improve the hydrogen utilization, but the nitrogen and water permeating from the cathode across the membrane to the anode may cover the three-phase reaction surface of the catalytic layer, resulting in a localized shortage of hydrogen at the anode, which leads to a reduction in the output power of the stack [6]. The recirculation mode uses a recirculation device to transport the hydrogen from the anode outlet to the anode inlet to participate in the reaction again, and since the anode exhaust gas contains water vapor generated by the electrochemical reaction and residual hydrogen, the recirculation process not only effectively improves the hydrogen utilization rate, but also has a certain humidification effect on the hydrogen at the anode inlet, which increases the effective utilization of the water of the fuel cell product, and this mode is the most widely used in fuel cell vehicles. This mode is the most widely used exhaust gas treatment method in fuel cell vehicles.

At present, the main two commonly used hydrogen circulation elements are the elicitor and the hydrogen circulation pump. Hydrogen circulation schemes mainly include: single elicitor, single hydrogen circulation pump, two-stage elicitor in parallel, elicitor and hydrogen circulation pump used in combination, and so on.

The research on the fuel cell system-based elicitor mainly focuses on the matching problem between the elicitor and the PEMFC system, including: the influence of the elicitor structure on the PEMFC system; the performance study of the elicitor under different operating parameters; the elicitor modelling method based on the PEMFC system, etc. Bao et al [7] established a dynamic model of the PEMFC system including elicitor and investigated the effect of the current on the performance of the elicitor by simulation, and the results showed that the elicitation ratio decreased abruptly when the current was instantaneously reduced and then increased rapidly. The effect of the current on the performance of the elicitor is investigated through simulation, and the results show that the elicitor ratio plummets to 0 and then rises back quickly when the current decreases instantaneously. Dadvar et al [8] investigated the correlation between the design

parameters of the stack and the design parameters of the elicitor, and by analyzing the effects of the activation area of the cell, the number of single cells, the diameter of the nozzle, and the diameter of the mixing chamber on the output efficiency of the fuel cell and the value of the current density corresponding to the maximum increment in efficiency, and based on this two dimensionless parameters, size ratio and diameter ratio, were proposed to establish a link between the design of the electric stack and the design of the inducer. MA et al [9] investigated the matching design problem between the inducer and the PEMFC system, quantified the actual boundary conditions of the inducer in the overall operating range, and established the design of the inducer including the cycle ratio, the hydrogen cycle ratio, the minimum current when the hydrogen cycle ratio is greater than 1.5, and the secondary current when the secondary current is wet and dry. A comprehensive elicitor performance evaluation system has been developed that includes four metrics: circulation ratio, hydrogen circulation ratio, minimum current at hydrogen circulation ratio greater than 1.5, and ratio of hydrogen circulation ratio at wetting and drying of secondary stream.

For hydrogen circulation pumps are currently divided into two main categories: volumetric and vane. Among them, Roots-type, claw-type, and scroll-type are volumetric pumps, while vortex-type are vane pumps [10]. Roots-type pumps have a double-rotor structure, with the main and driven shafts parallel, and compress the gaseous medium through the rotary motion of the cam rotor. Roots-type pumps have higher pressure rise, smooth operation, low noise, low vibration, and no internal oil lubrication is required, which avoids contamination of the system by oil vapor. Claw type hydrogen circulation pump has two claw rotors rotating in opposite directions, the two rotors will not contact each other, and there is a very small gap between the rotor and the chamber shell. Claw pumps are stable, but have poor sealing, vibration and noise. Vortex pumps are mainly composed of fixed and driven two vortex discs. When driving the scroll disc rotation, the gas from the outer edge of the inhalation and compression between the two scroll discs and transported to the center of the scroll disc. Because of the small clearance within the scroll pump, there is less gas leakage, lower vibration and noise, but the pressure rise and flow rate of this type of pump is also smaller. The vane pump converts the mechanical energy of the vane into the kinetic energy of the fluid through the rotation of the impeller. Vane pumps are simple in structure, smaller in size and produce less energy consumption. The energy consumption and noise problems of hydrogen circulation pumps are the main reasons for limiting their use. He Lingxuan [11] established an energy consumption model for hydrogen circulation pumps for research, and the results showed that the energy loss of hydrogen circulation pumps is about 23.3% of the loss of fuel cell appurtenant equipment.

The combination of the pilot and the hydrogen circulation pump as the hydrogen circulation device of the PEMFC system can ensure the circulation demand at a very low power and can play a better performance for the scenario of frequent load change. When the fuel cell is in the low power zone, the ejector performance is not good, then the hydrogen circulation pump is activated to circulate hydrogen, when in the high-power zone, the ejector as the main circulation device can meet the demand. This scheme not only avoids the problem of poor performance of the pilot in the low power region, but also reduces the power consumed by the hydrogen circulation pump.

The combination of the pilot and the hydrogen circulation pump as the hydrogen circulation device of the PEMFC system can ensure the circulation demand at a very low power and can play a better performance for the scenario of frequent load change. When the fuel cell is in the low power zone, the ejector performance is not good, then the hydrogen circulation pump is activated to circulate hydrogen, when in the high-power zone, the ejector as the main circulation device can meet the demand. This scheme not only avoids the problem of poor performance of the ejector in the low-power region, but also reduces the power consumed by the hydrogen circulation pump. However, at this stage, the matching control strategy of the ejector and hydrogen circulation pump under on-board operating conditions needs to be further investigated. Therefore, based on the automotive fuel cell system, this paper designs a hydrogen recirculation system to meet its operational requirements. Finally, the matching control strategy of the hydrogen circulation subsystem is established for different operating conditions of the vehicle fuel cell.

2 DEMAND ALLOCATION FOR EJECTORS AND HYDROGEN CIRCULATION PUMPS

2.1 Fuel Cell Hydrogen Cycle System Structure

The design of hydrogen circulation system for on-board fuel cell puts higher requirements on the circulation volume under the whole power range and power loss. The traditional single ejector or hydrogen circulation pump can no longer meet the demand of the electric stack cycle, so this paper improves the hydrogen system structure of the on-board fuel cell system, adopting a combination of ejectors and hydrogen circulation pumps, the structural principle of the hydrogen circulation system is shown in Figure 1.

Hydrogen released from the high-pressure hydrogen storage cylinder enters the inlet of the hydrogen recycling device, i.e. the primary inflow port of the ejector, after the pressure is adjusted by the pressure reducing valve and the proportional valve; after the anode tail gas passes through the water separator, a part of the gas and the liquid water are discharged through the drainage and exhaust solenoid valve periodically, and a part of the gas enters the secondary inflow port of the ejector or the hydrogen recycling pump, and then it mixes with the new hydrogen and enters into the fuel cell to participate in the electrochemical reaction again. As the hydrogen demand of the stack is different under different working conditions, the corresponding flow resistance is also different. Therefore, the design of hydrogen circulation system should meet the demand of hydrogen circulation flow on the one hand, and the pressure demand on the other hand.

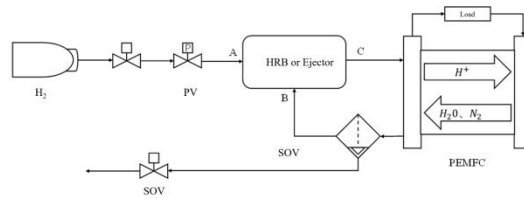


Figure 1 Schematic Diagram of Fuel Cell Hydrogen System

2.2 Hydrogen Cycle Flow Requirements

In fuel cell systems, in order to ensure that the hydrogen supply is sufficient and there is no shortage of reactants when the load changes suddenly, the hydrogen supplied to the anode of the fuel cell is generally in excess, and the actual flow rate of hydrogen is about 1.1 to 1.5 times of the theoretical flow rate [12, 13]. The volume flow rates of fresh hydrogen and circulating hydrogen as well as the molar fractions (volume fractions) of each component in the circulating gas in the 100kW fuel cell system at various currents were measured as shown in Table 1.

Table 1 Hydrogen Cycle Flow Requirements for Electric Stack

Current (A)	Power (kW)	H2_in (nlpm)	H2_re (nlpm)	Flow_re (nlpm)	mole fraction		
					H2	N2	Vapor
50	13.3	103.9	177	377.8	0.468	0.415	0.116
100	25.2	207.9	257.2	510.1	0.504	0.365	0.131
150	36.4	311.9	316.4	595.4	0.531	0.326	0.143
200	47	415.8	350.1	639.2	0.548	0.308	0.144
225	52.2	467.8	364.1	653	0.558	0.3	0.142
240	55.3	499.1	376.9	664.8	0.567	0.292	0.141
300	66.9	623.9	511.8	864.1	0.592	0.272	0.135
350	76.5	728	582.4	960.8	0.606	0.267	0.127
400	85.7	832	693.8	1132.3	0.613	0.265	0.122
450	95.2	936.1	821.7	1328.8	0.618	0.265	0.117
500	103.5	1040.2	986.1	1594.1	0.619	0.265	0.116

The volumetric flow rates of new and circulating hydrogen in Table 1 were converted to mass flow rates using the following conversion equations:

$$Q_m = \frac{Q_v \times \rho}{60} \quad (1)$$

Where Q_m is the mass flow rate of gas (g/s), Q_v is the volume flow rate of gas (lpm), and ρ is the density of gas (kg/m³). According to the test data of the actual gas volume flow rate and mass flow rate of the fuel cell system, the density of hydrogen in the converted high-pressure hydrogen and circulating gas is about 0.083 kg/m³.

The mass flow rate of the circulating gas is calculated by the following formula:

$$Q_R = \frac{Q_{R,H_2}}{w_{H_2}} \quad (2)$$

Where Q_R is the mass flow rate of circulating gas (g/s), Q_{R,H_2} is the mass flow rate of hydrogen in the circulating gas, and w_{H_2} is the mass fraction of hydrogen in the circulating gas. w_{H_2} is obtained based on the mole fraction, which is given by the following formula:

$$w_{H_2} = \frac{x_{H_2} M_{H_2}}{x_{H_2} M_{H_2} + x_{N_2} M_{N_2} + x_{H_2O} M_{H_2O}} \quad (3)$$

where x_{H_2} , x_{N_2} , x_{H_2O} are the molar fractions of hydrogen, nitrogen, and water vapor in the recycle gas, and M_{H_2} , M_{N_2} , M_{H_2O} are the molar masses of hydrogen, nitrogen, and water vapour (2 g/mol, 28 g/mol, and 18 g/mol), respectively.

The gas mass flow rate and cycle ratio demand at each current of the fuel cell were calculated according to Eqs. (1) to (3) as shown in Table 2.

Table 2 Cycle Ratio Requirements for Electric Stacks

Current(A)	H2_in(g/s)	H2_re(g/s)	Ration
50	0.14	3.75	26.79

100	0.29	4.85	16.72
150	0.43	5.29	12.3
200	0.58	5.39	9.29
225	0.65	5.41	8.32
240	0.69	5.43	7.87
300	0.86	6.73	7.83
350	1.01	7.33	7.26
400	1.15	8.49	7.38
450	1.29	9.93	7.7
500	1.44	11.8	8.19

2.3 Hydrogen Cycle Pressure Requirements

In fuel cell systems, the anode inlet and outlet usually have a certain pressure, which is crucial for promoting electrochemical reactions, and the pressure drop at the anode under different currents can show significant differences. Therefore, the design of the hydrogen circulation system must consider the pressure characteristics of the fuel cell under various operating conditions. The inlet and outlet pressures of the stack under each current were obtained by testing as shown in Table 3.

Table 3 Hydrogen Cycle Pressure Requirements

Current(A)	P_in(kPa)	P_out(kPa)	P_in/P_out
50	132	126	1.048
100	138	130	1.062
150	142	133	1.068
200	148	138	1.072
225	152	142	1.07
240	155	145	1.069
300	166	155	1.071
350	181	169	1.071
400	197	184	1.071
450	220	206	1.068
500	237	221	1.072

2.4 Hydrogen Circulation Pump Selection

Since the cycling performance of the elicitor is poor at small currents and the power consumed by the circulation pump increases with the increase of the rotational speed, when the combination of the elicitor and the hydrogen circulation pump is selected as the hydrogen cycling device, the elicitor is made to exert its performance at large currents while the circulation pump mainly works at small currents. Based on this, the current interval of 0~500A of the fuel cell is initially divided according to 60% of the maximum current, i.e., divided into two working intervals of 0~300A and 300~500A. When making the selection of hydrogen circulation pump, the performance of hydrogen circulation pump should at least meet the circulation demand under 0~300A. According to Tables 1 and 3, the hydrogen circulation system should be able to meet the flow rate of 377.8~864.1 lpm at a pressure ratio of less than 1.072.

Based on the requirement analysis, a hydrogen circulation pump with MAP as shown in Figure 2 was selected as follows. A single pump can circulate gas flow between 185.6~1150slpm at a pressure ratio of 1.072, which can cover the circulating demand under 0~300A.

Since the inducer performs better at high currents, it is expected that the circulating pump will only need to maintain relatively low speed operation at high currents to meet the demand. The maximum demand flow rate of the stack is 1594.1nlpm, and further considering the reliability and fault tolerance of the system, when designing the ejector, the maximum circulating gas flow rate of the ejector and the hydrogen circulation pump at a certain pressure ratio should reach 1650nlpm. at the same time, the ejector needs to overcome the anode flow resistance of the fuel cell, and the pressure ratio has to be greater than 1.072. Therefore, the secondary flow rate at a pressure ratio of 1.072 and a secondary flow rate of 1650nlpm is taken as the design target of the ejector. 1650nlpm secondary flow rate as the design goal of the pilot ejector.

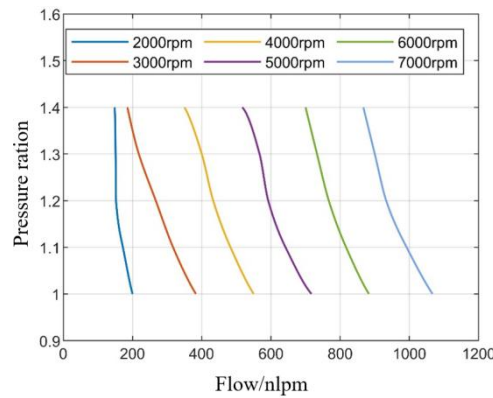


Figure 2 MAP Diagram of a Certain Type of Hydrogen Circulation Pump

3 MODELLING OF KEY COMPONENTS

The hydrogen recirculation system designed in this paper has two key components, which are the inducer and the hydrogen recirculation pump. In order to study the arrangement scheme and matching control strategy of the ejector and the hydrogen circulation pump, one-dimensional modelling of the ejector and the hydrogen circulation pump is carried out.

3.1 Ejector

The one-dimensional modelling of the ejector is carried out based on the one-dimensional gas dynamics of compressible gases, and the flow characteristics of the fluid are solved by applying the equations of conservation of mass, conservation of momentum, and conservation of energy. When the structural parameters of the ejector are known, its performance can be evaluated by the one-dimensional model, which calculates the flow characteristics of the fluid in the primary flow inlet, the inhalation chamber, the secondary flow inlet, the mixing chamber, and the diffusion chamber in turn according to the axial position, and then calculates the ejection ratio. In addition, when designing the ejector, the ejector 1D model is also capable of outputting structural parameters from the performance parameters.

The one-dimensional model of the pilot is based on the following assumptions:

- (1) The gases are all ideal gases;
- (2) The velocity of the primary flow is uniform in the radial direction;
- (3) The internal walls of the inducer are adiabatic;
- (4) The calculation of friction losses is isentropic.

3.1.1 Primary flow

When the primary flow flows from the inlet to the nozzle outlet, the flow velocity and pressure characteristics of the gas vary greatly according to Bernoulli's principle. The flow velocity at the outlet of the constricted nozzle is classified into sonic and subsonic flow according to the critical value v_{cr} of the ratio of the pressure of the secondary flow to the primary flow, P_s/P_p , as shown in Figure 3. When the pressure ratio is less than the critical value v_{cr} , the fluid flow rate does not follow the pressure ratio, while when the pressure ratio is greater than the critical value v_{cr} , the fluid flow rate decreases with the increase of the pressure ratio.

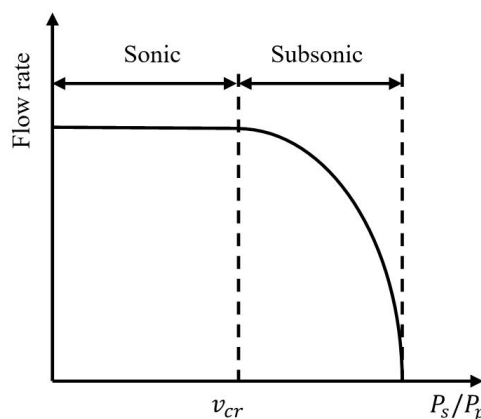


Figure 3 Flow Characteristics at the Outlet of a Shrinkage Nozzle

In this case, the formula for calculating the critical value is as follows:

$$v_{cr} = \left(\frac{2}{k_p + 1} \right)^{\frac{k_p}{k_p - 1}} \quad (4)$$

Where, k_p is the gas adiabatic index, which is taken as $k_p=1.41$ since the primary stream of the inducer is pure hydrogen, which is a diatomic molecule.

The flow characteristics from the primary flow inlet to the nozzle outlet are calculated according to the isentropic flow law to obtain the mass flow rate and Mach number at the primary flow nozzle outlet. Where Mach number is the ratio of the fluid velocity to the speed of sound in its surrounding medium.

1) For $\frac{P_s}{P_p} < v_{cr}$, the fluid is in acoustic flow, and the primary flow rate and outlet Mach number are:

$$m'_{p,0} = A_t P_{p,0} \sqrt{\frac{\psi_{p,0} k_p}{R_{g,p} T_{p,0}}} \left(\frac{2}{k_p + 1} \right)^{\frac{k_p + 1}{2(k_p - 1)}} \quad (5)$$

$$M_t = 1 \quad (6)$$

2) For $\frac{P_s}{P_p} \geq v_{cr}$, the fluid is in subsonic flow, and the primary flow rate and outlet Mach number are:

$$m'_{p,0} = A_t P_{p,0} \sqrt{\frac{2 \psi_p k_p \left[\left(\frac{P_{s,0}}{P_{p,0}} \right)^{\frac{2}{k_p}} - \left(\frac{P_{s,0}}{P_{p,0}} \right)^{\frac{k_p + 1}{k_p}} \right]}{(k_p - 1) R_{g,p} T_{p,0}}} \quad (7)$$

$$M_t = \sqrt{\frac{2 \left[1 - \left(\frac{P_{s,0}}{P_{p,0}} \right)^{\frac{k_p - 1}{k_p}} \right]}{k_p - 1}} \quad (8)$$

where subscripts 0 and t denote the fluid properties at the inlet of the primary or secondary flow and at the outlet of the nozzle, respectively, subscripts p and s denote the primary and secondary flow, respectively, m is the mass flow rate (kg/s), M is the Mach number, A is the cross-sectional area (m²), P is the pressure (Pa), ψ is the isentropic coefficient considering friction losses, R_g is the gas constant (J/(kg·K)), and T is the temperature (K).

In order to improve the accuracy of the model, the primary flow rate was corrected according to the experimental data, and the corrected formula was:

$$m_{p,0} = \begin{cases} 1.6m'_{p,0} & P_{p,0} \leq 100\text{kPa} \\ 1.2m'_{p,0} & 100\text{kPa} < P_{p,0} \leq 300\text{kPa} \\ m'_{p,0} & 300\text{kPa} < P_{p,0} \leq 600\text{kPa} \\ 0.92m'_{p,0} & 600\text{kPa} < P_{p,0} \leq 800\text{kPa} \\ 0.91m'_{p,0} & P_{p,0} > 800\text{kPa} \end{cases} \quad (9)$$

3.1.2 Inhalation chamber

The one-dimensional model of a conventional ejector assumes that the pressure in the inhalation chamber is equal to the secondary inflow pressure, however, the pressure of the primary flow from the nozzle outlet to the inhalation chamber decreases, and the pressure difference with the secondary flow pressure causes the secondary flow to be sucked in. In order to improve the accuracy of the model, the formula for the pressure in the inhalation chamber in this paper adopts the modified formula in the literature [14]:

$$P_{p,2} = \begin{cases} 0.957P_{s,0} & (P_{p,0} \leq 125\text{kPa}) \\ 0.895P_{s,0} & (125\text{kPa} < P_{p,0} \leq 150\text{kPa}) \\ 0.845P_{s,0} & (150\text{kPa} < P_{p,0} \leq 175\text{kPa}) \\ 0.795P_{s,0} & (175\text{kPa} < P_{p,0} \leq 200\text{kPa}) \\ 0.690P_{s,0} & (200\text{kPa} < P_{p,0} \leq 250\text{kPa}) \\ 0.570P_{s,0} & (250\text{kPa} < P_{p,0} \leq 300\text{kPa}) \\ 0.470P_{s,0} & (300\text{kPa} < P_{p,0} \leq 400\text{kPa}) \\ 0.400P_{s,0} & (400\text{kPa} > P_{p,0}) \end{cases} \quad (10)$$

where the subscript 2 indicates the flow characteristics of the fluid at the inlet of the mixing chamber.

Based on this, the flow characteristics of the primary flow at cross-section 2 are calculated by assuming that the primary flow has a constant velocity over a certain range at cross-section 2 according to the isentropic flow law and the law of conservation of energy:

$$M_{p,2} = \sqrt{\frac{2 \left(\frac{P_{p,0}}{P_{s,0}} \right)^{\frac{k_p - 1}{k_p}} - 2}{k_p - 1}} \quad (11)$$

$$T_{p,2} = \frac{T_{p,0}}{1 + \frac{1}{2}(k_p - 1)M_{p,2}^2} \quad (12)$$

$$V_{p,2} = M_{p,2} \sqrt{k_p R_{g,p} T_{p,2}} \quad (13)$$

$$D_{p,2} = \frac{D_t}{\Psi_{\text{exp}}} \sqrt{\frac{M_t}{M_{p,2}}} \left[\frac{2 + (k_p - 1)M_{p,2}^2}{2 + (k_p - 1)M_t^2} \right]^{\frac{k_p + 1}{4(k_p - 1)}} \quad (14)$$

Where V denotes the fluid velocity (m/s), D is the cross-sectional diameter (m), and Ψ_{exp} is the coefficient of friction loss at the beginning of mixing of the primary and secondary flows.

3.1.3 Secondary flow

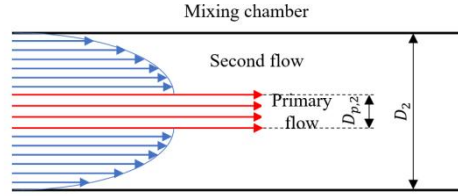


Figure 4 Schematic of the Velocity Distribution at the Entrance to the Mixing Chamber of the Inducer

The secondary flow is distributed outside the primary flow as it flows to the inlet 4 of the mixing chamber, and its velocity is not uniform in the radial direction, as shown in Figure 4. The primary flow is within $D_{p,2}$, while the secondary flow is predominant within the mixing chamber beyond $D_{p,2}$. The velocity function at the inlet of the mixing chamber with respect to primary and secondary flow is defined as [15]:

$$v_r = \begin{cases} V_{p,2} & , 0 \leq r \leq R_{p,2} \\ V_{p,2} \left(\frac{R_2 - r}{R_2} \right)^{\frac{1}{n_v}} & , R_{p,2} \leq r \leq R_2 \end{cases} \quad (15)$$

Where R is the cross-section radius (m), n_v is the velocity coefficient, a parameter related to the nozzle outlet diameter, mixing chamber diameter, primary and secondary flow pressure, which can be calculated by empirical formula:

$$n_v = 1.393 \times 10^{-4} e^{\frac{\beta_p}{0.05}} + 0.456 \beta_D + 0.1668 \quad (16)$$

Where β_p is a pressure parameter and β_D is a structural parameter calculated as follows:

$$\beta_p = \frac{p_{s,0}^{0.8}}{p_p^{1.1}} \quad (17)$$

$$\beta_D = \frac{D_2}{D_t} \quad (18)$$

Based on the velocity function, the secondary flow rate at the inlet of the mixing chamber is calculated by the following equation:

$$m_{s,2} = \int_{R_{p,2}}^{R_2} \overline{\rho}_{s,0} v_r dA_2 \quad (19)$$

where $\overline{\rho}_{s,0}$ is the average density of the secondary stream (kg/m³), which can be calculated from the ideal gas equation of state. The mass fraction of nitrogen and water vapour is first calculated from the mass fraction of hydrogen in the secondary stream:

$$w_{N_2} = \frac{x_{N_2} M_{N_2}}{x_{H_2} M_{H_2} + x_{N_2} M_{N_2} + x_{H_2O} M_{H_2O}} \quad (20)$$

$$w_{H_2O} = \frac{x_{H_2O} M_{H_2O}}{x_{H_2} M_{H_2} + x_{N_2} M_{N_2} + x_{H_2O} M_{H_2O}} \quad (21)$$

Then, the average density of the secondary flow gas is calculated from the ideal gas equation of state as follows:

$$\overline{\rho}_{s,0} = \frac{p_{s,0}}{R_u T_{s,0}} (M_{H_2} \cdot w_{N_2} + M_{N_2} \cdot w_{N_2} + M_{H_2O} \cdot w_{H_2O}) \quad (22)$$

where R_u is the molar gas constant (J/(mol·K)).

Mass flow rate of the secondary flow at the inlet of the mixing chamber:

$$m_{s,2} = 2\pi V_{p,2} \overline{\rho}_{s,0} \left[\frac{n_v R_2^2}{n_v + 1} \left(1 - \frac{R_{p,2}}{R_2} \right)^{\frac{n_v + 1}{n_v}} - \frac{n_v R_2^2}{2n_v + 1} \left(1 - \frac{R_{p,2}}{R_2} \right)^{\frac{2n_v + 1}{n_v}} \right] \quad (23)$$

At this point, both the primary and secondary flow rates are known, and the priming ratio can be calculated by the following equation:

$$\lambda = \frac{m_{s,2}}{m_{p,0}} \quad (24)$$

Calculate the average velocity of the secondary flow at the inlet of the mixing chamber:

$$V_{s,2} = \frac{m_{s,2}}{\overline{\rho}_{s,0} A_{s,2}} = \frac{2\pi V_{p,2}}{A_{s,2}} \left[\frac{n_v R_2^2}{n_v + 1} \left(1 - \frac{R_{p,2}}{R_2} \right)^{\frac{n_v + 1}{n_v}} - \frac{n_v R_2^2}{2n_v + 1} \left(1 - \frac{R_{p,2}}{R_2} \right)^{\frac{2n_v + 1}{n_v}} \right] \quad (25)$$

where $A_{s,2}$ is the cross-sectional area of the flow region of the secondary flow at the inlet of the mixing chamber, calculated as:

$$A_{s,2} = \pi(R_2^2 - R_{p,2}^2) \quad (26)$$

The Mach number, pressure and temperature of the secondary flow at the inlet of the mixing chamber are calculated by the following equation:

$$M_{s,2} = \frac{V_{s,2}}{\sqrt{k_s R_{g,s} T_s}} \quad (27)$$

$$P_{s,2} = \frac{P_{s,0}}{\left(1 + \frac{k_s - 1}{2} M_{s,2}^2\right)^{\frac{k_s}{k_s - 1}}} \quad (28)$$

$$T_{s,2} = \frac{T_s}{1 + \frac{1}{2}(k_s - 1)M_{s,2}^2} \quad (29)$$

where k_s is the gas adiabatic index of the secondary flow of the inducer. Since the model assumes that the secondary flow is an ideal gas, k_s is equal to the specific heat ratio, which is calculated as follows [16]:

$$k_s = k_{H_2} w_{N_2} + k_{N_2} w_{N_2} + k_{H_2O} w_{N_2O} \quad (30)$$

Where k_{H_2} , k_{N_2} , k_{H_2O} are the specific heat ratios of hydrogen, nitrogen, and water vapour, respectively.

3.1.4 Mixing chambers

In the process of mixing primary and secondary flow in the mixing chamber, the fluid velocity, temperature and pressure can be calculated by the equations of conservation of momentum, conservation of energy and conservation of mass, and the fluid characteristics at the outlet of the mixing chamber are calculated as follows:

$$V_3 = \frac{\Psi_{mix}(m_{p,0}V_{p,2} + m_{s,2}V_{s,2})}{m_{p,0} + m_{s,2}} \quad (31)$$

$$T_3 = \frac{1}{C_{p,mix}} \left[\frac{m_{p,0} \left(C_{p,p,2} T_{p,2} + \frac{V_{p,2}^2}{2} \right) + m_{s,2} \left(C_{p,s,2} T_{s,2} + \frac{V_{s,2}^2}{2} \right)}{m_{p,0} + m_{s,2}} - \frac{V_3^2}{2} \right] \quad (32)$$

$$P_3 = \frac{(m_{p,0} + m_{s,2}) \cdot R_{g,mix} T_3}{V_3 A_3} \quad (33)$$

Where Ψ_{mix} is the friction loss coefficient for mixing primary and secondary flows in the mixing chamber, $C_{p,mix}$ is the constant pressure specific heat capacity of the gas mixture (J/(kg·K)), and $R_{g,mix}$ is the gas constant of the gas mixture (J/(kg·K)). $C_{p,mix}$ is calculated using the following formula:

$$C_{p,mix} = C_{p,H_2} w'_{H_2} + C_{p,N_2} w'_{N_2} + C_{p,H_2O} w'_{H_2O} \quad (34)$$

where C_{p,H_2} , C_{p,N_2} , C_{p,H_2O} are the constant pressure specific heat capacity of hydrogen, nitrogen, and water vapour (J/(kg·K)), respectively, and w'_{H_2} , w'_{N_2} , w'_{H_2O} are the mass fractions of hydrogen, nitrogen, and water vapour in the gases mixed in the primary and secondary streams, respectively.

The Mach number of the gas at the exit of the mixing chamber is:

$$M_3 = \frac{V_3}{\sqrt{k_3 R_{g,mix} T_3}} \quad (35)$$

where k_3 is the gas adiabatic index of the secondary flow of the inducer, calculated in the same way as k_s .

3.1.5 Diffusion chambers

The flow of the gas mixture in the diffusion chamber is a decreasing velocity and increasing pressure process, and the pressure at the outlet of the inducer is calculated by the isentropic flow law:

$$P_4 = P_3 \left[1 + \frac{1}{2} (k_3 - 1) M_3^2 \right]^{\frac{k_3}{k_3 - 1}} \quad (36)$$

The temperature at the exit of the ejector is calculated according to the law of conservation of energy from the inlet to the outlet of the ejector:

$$T_4 = \frac{m_p C_{p,H_2} T_p + (m_{s,H_2} C_{p,H_2} + m_{s,N_2} C_{p,N_2} + m_{s,H_2O} C_{p,H_2O}) T_s - E_{loss}}{(m_p + m_{s,H_2} + m_{s,N_2} + m_{s,H_2O}) C_{p,mix}} \quad (37)$$

Where E_{loss} is the energy loss (J/s) of primary and secondary flow, which is calculated as follows:

$$E_{loss} = \frac{1}{2} (1 - \Psi_p) m_p V_{p,2}^2 + \frac{1}{2} (1 - \Psi_s) m_s V_{s,2}^2 \quad (38)$$

3.1.6 Validation of the elicitor model

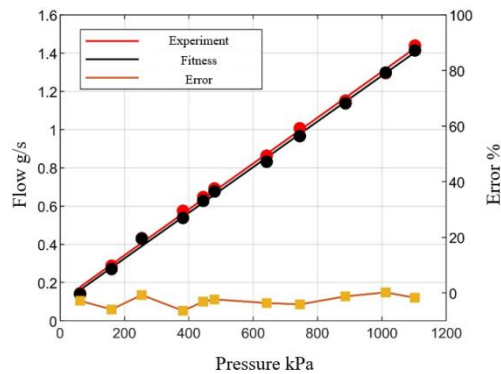


Figure 5 Validation of the Model of the Introducer

In order to verify the accuracy of the pilot model, this paper simulates the pilot and derives the experimental data under the same boundary conditions as the simulation. By fitting the experimental and simulation data and calculating the relative error, the results are shown in Figure 5. It can be seen that the simulation and experimental results of the primary flow rate have the same trend, and the relative error of the two is between -6.4% and 0.2%, which can be concluded that the pilot model established in this paper has high accuracy and can be used for subsequent research.

3.2 Hydrogen Circulation Pumps

The one-dimensional model of the hydrogen circulation pump consists of three sub-models, namely the flow-pressure ratio relationship model, the torque and angular velocity model of the drive motor, and the PID control model of the motor.

3.2.1 Flow-pressure ratio model

The flow-pressure ratio relationship model of the hydrogen circulating pump is based on the test data and established by the method of checking the table. The model takes the component partial pressures of the inlet and outlet gases of the circulating pump and the rotational speed of the circulating pump as input parameters, and can output the circulating pump outlet gas flow rate and the amount of power loss.

3.2.2 Torque and angular velocity modelling of the drive motor

The torque and angular velocity model of the hydrogen circulating pump motor is based on the following equations:

$$\frac{d\omega_l}{dt} = \frac{1}{J_l} (\tau_m - \tau_l) \quad (39)$$

where the subscripts l and m denote the load and drive motor parameters, respectively, ω is the angular velocity of the rotor (rad/s); t is the time (s); J is the rotational moment of inertia of the rotor of the hydrogen circulating pump ($\text{kg}\cdot\text{m}^2$); and τ is the torque ($\text{N}\cdot\text{m}$). Among them, the driving torque and load torque are calculated by the following equation:

$$\tau_m = \eta_m \frac{\kappa_t}{R_m} (u_m - \kappa_v \omega_l) \quad (40)$$

$$\tau_l = \frac{P_l}{\omega_l} \quad (41)$$

Where η is the efficiency; R_m is the motor resistance (Ω); κ_t , κ_v denote the torque constant and voltage constant of the motor, respectively; u_m denotes the motor control voltage (V). P_l is the power consumed by the hydrogen circulation pump, which is calculated by the following formula:

$$P_l = C_{p_{in}} \frac{T_{in}}{\eta_l} \left[\left(\frac{P_{out}}{P_{in}} \right)^{\frac{k_{in}-1}{k_{in}}} - 1 \right] m_{in} \quad (42)$$

Where $C_{p_{in}}$ is the constant-pressure specific heat capacity of the inlet gas of the circulating pump ($\text{J}/(\text{kg}\cdot\text{K})$); T_{in} is the temperature of the inlet gas (K); P_{out} , P_{in} represent the pressures of the outlet and the inlet, respectively (Pa); k_{in} is the specific heat ratio of the inlet gas; and m_{in} is the mass flow rate of the inlet gas (kg/s).

The formula for calculating the outlet gas temperature of a hydrogen circulation pump is as follows:

$$T_{out} = T_{in} + \frac{T_{in}}{\eta_l} \left[\left(\frac{P_{out}}{P_{in}} \right)^{\frac{k_{in}-1}{k_{in}}} - 1 \right] \quad (43)$$

Where T_{out} is the circulating pump outlet temperature (K).

3.2.3 PID control model for motors

In the hydrogen circulation system, the proportional valve serves to control the inlet pressure, and the speed of the hydrogen circulation pump has a significant effect on the circulation system flow and pressure, therefore, precise control of the circulation pump is required.

When the circulating pump speed is stable, its control voltage is calculated by means of the angular velocity, which is calculated as follows:

$$u_m = \kappa_v \omega_l + \frac{P_l R_m}{\omega_l \eta_m \kappa_t} \quad (44)$$

Calculate the amount of deviation of the actual angular velocity from the target angular velocity:

$$e(t) = \omega_{set} - \omega_l \quad (45)$$

where ω_{set} is the target angular velocity (rad/s) and $e(t)$ is the angular velocity deviation (rad/s).

The deviation amount of angular velocity, $e(t)$, is used as an input to the PID to calculate the control voltage after control:

$$u(t) = K_p e(t) + K_I \int e(t) dt + K_D \frac{de(t)}{dt} \quad (46)$$

$$u_{m,new} = u_m + u(t) \quad (47)$$

Where, K_p , K_I , K_D represent proportional, integral and differential coefficients, respectively; $u(t)$ is the change of control voltage (V); $u_{m,new}$ is the control voltage (V) after PID control.

Set the PID coefficients as $K_p = K_I = K_D = 1$, the rotational speed of circulating pump is correspondingly as shown in Fig. 2.7(a), and the deviation from the target rotational speed is large. Adjust the PID parameters to $K_p = 150$, $K_I = 5$, $K_D = 100$, the control effect is shown in Figure 6, the circulating pump speed to reach the target speed of the time is greatly shortened, to achieve efficient control.

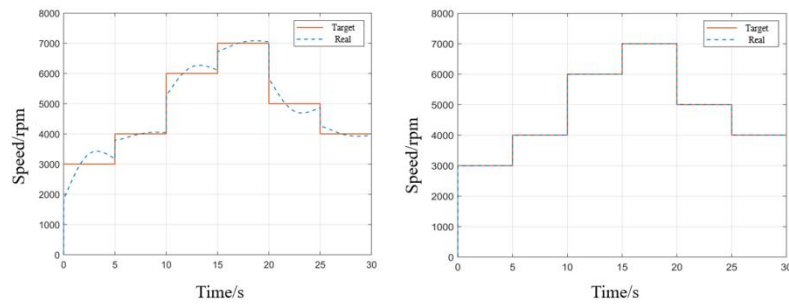


Figure 6 PID Control Effect of Hydrogen Circulating Pump

3.2.4 Hydrogen circulation pump model validation

In order to verify the accuracy of the hydrogen circulating pump model, a set of data was randomly selected for simulation and compared with the experimental results. During the simulation process, the inlet pressure of the circulating pump is set at 100 kPa and the rotational speed is 6000 rpm, and the comparison between the obtained volume flow rate simulation data and the experimental data is shown in Figure 7. The maximum relative error between the simulation data and experimental data is 0.82%, and it can be concluded that the accuracy of the established circulation pump model is high.

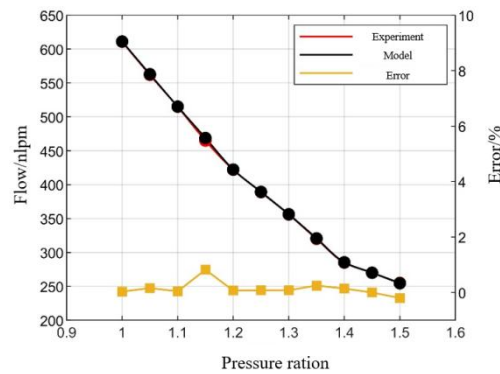


Figure 7 Hydrogen Circulation Pump Model Validation

4 MATCHING CONTROL STRATEGY FOR HYDROGEN CIRCULATION SYSTEM

4.1 Ejector in Series with a Hydrogen Circulation Pump

When the ejector is connected in series with the hydrogen circulation pump, the gas at the outlet of the circulation pump is used as the secondary flow of the ejector, and the sum of the pressures of the ejector and the circulation pump is equal to the anode flow resistance of the fuel cell.

In the state where the ejector and the hydrogen circulation pump are connected in series, they share the responsibility of overcoming the flow resistance of the fuel cell anode. Based on the pressure requirement of the fuel cell anode, the minimum pressure rise of the hydrogen circulation pump can be calculated by the following equation:

$$\Delta P_{pump} = \Delta P_{anode} - \Delta P_{ejector} \quad (48)$$

Since the ejector and the hydrogen circulation pump are connected in series, they share the pressure of the anode, and in order to ensure the circulating flow of the hydrogen circulation system, the hydrogen circulation pump has to be operated continuously over the entire current range. Therefore, it is necessary to carry out a reasonable pressure distribution between the ejector and the hydrogen circulation pump. On the one hand, when the circulating pump speed is certain, the pressure rise of the circulating pump should be made lower to ensure more circulating gas volume, so that the elicitor should share as much pressure as possible; on the other hand, under the premise of satisfying the circulating volume, the circulating pump speed should be set smaller in order to reduce the power consumption (Figure 8).

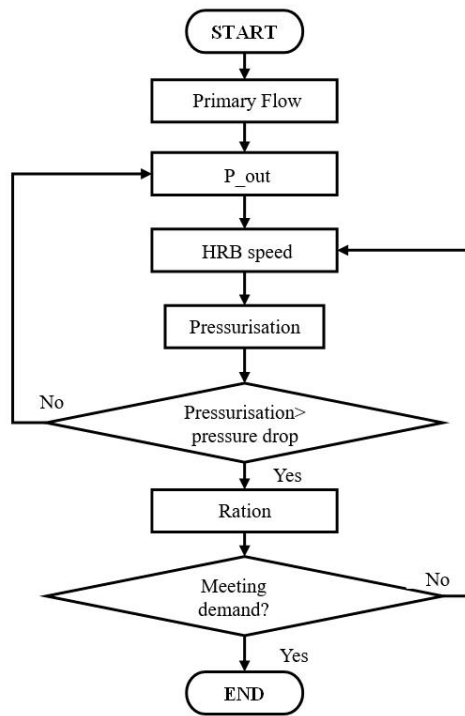


Figure 8 Flow of Formulation of Matching Strategy in the Cascade Mode of Inducer and Hydrogen Recirculation Pump

4.2 Parallel Connection of the Ejector and the Hydrogen Circulation Pump

The characteristics of pilot and hydrogen circulating pump in parallel are ‘equal pressure rise, divided flow’, that is, the pressure rise of both is equal to the anode pressure drop, and the circulating flow is shared by both.

In the current range of 100A and below, the injector can not overcome the anode flow resistance of the fuel cell, and then the phenomenon of secondary flow reversal may occur, which may lead to a significant reduction in the cycle efficiency, and even affect the normal operation of the fuel cell. In order to avoid the occurrence of reverse flow phenomenon, this paper adds a check valve in the secondary flow gas flow path of the ejector. After adding the check valve, the backflow phenomenon of the secondary flow was alleviated, but it also resulted in almost no circulating gas entering the elicitor. Therefore, the hydrogen recirculation pump is considered to work alone in the current range of 100 A and below (See Figure 9).

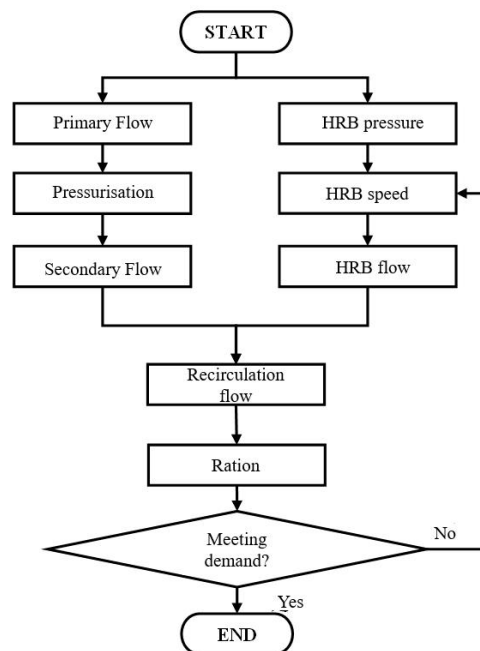


Figure 9 Flow of Formulation of Matching Strategy in Parallel Mode of Inducer and Hydrogen Circulating Pumps

The matching strategies of the pilot and the hydrogen circulation pump under two arrangement schemes, series and parallel, have been developed in the previous section, and both arrangement schemes can meet the hydrogen circulation demand of the fuel cell system. In the series and parallel schemes, the speed of the hydrogen circulation pump at each current point is shown in Table 4, and its average power consumption is 0.42kW and 0.39kW, respectively, and the power consumption under the parallel scheme is 7.1% smaller than that of the series scheme. Therefore, it was finally determined that the arrangement scheme of the ejector and the hydrogen circulation pump was parallel connection.

Table 4 Comparison of Hydrogen Circulation Pump Speed in Different Hydrogen Circulation Modes

Current (A)	Series HRB speed (rpm)	Parallel HRB speed (rpm)
50	3000	5000
100	4000	5000
150	4000	5000
200	4000	5000
225	4000	5000
240	4000	3000
300	5000	3000
350	5000	3000
400	6000	3000
450	6000	3000
500	6000	3000

5 CONCLUSION

In this paper, the optimized design of the hydrogen circulation system is carried out for automotive fuel cell systems. For the fuel cell electric stack, the flow rate and pressure demand of the circulation device design are analysed, the demand allocation for the elicitor and the hydrogen circulation pump is carried out, and the model of the hydrogen circulation pump is determined.

According to the one-dimensional design theory of the ejector, a one-dimensional model of the ejector is established; based on the experimental data of the hydrogen circulation pump, its one-dimensional model is established. Based on the experimental data, the one-dimensional model is verified to ensure the accuracy of the model.

Finally, based on the one-dimensional model of the ejector and the hydrogen circulation pump, the pressure rise or flow rate of the ejector and the hydrogen circulation pump in series and parallel are calculated, and the matching strategy is formulated. By comparing the power consumption under the series and parallel schemes, it is found that the average power consumption under the parallel scheme is 7.1% smaller than that of the series, so the parallel scheme is selected.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

FUNDING

This work was supported by China National Railway Group Co., Ltd. Unveiled Its Flagship Project (grant number: N2022J016-A) and Changzhou City's "Unveiling the List and Leading the Way" Science and Technology Research Project (grant number: 2023-Z-GKB-JS-0009).

REFERENCES

- [1] Mohsen Kandi Dayeni, Mehdi Soleymani. Intelligent energy management of a fuel cell vehicle based on traffic condition recognition. *Clean Technologies and Environmental Policy*, 2016, 18(6): 1945-1960.
- [2] Zhu Mengqian, Xie Xu, Wu Kangcheng, et al. Experimental investigation of the effect of membrane water content on PEM fuel cell cold start. *Energy Procedia*, 2019, 158: 1724-1729.
- [3] Kairui Dong, Guangbin Liu. A review of hydrogen recirculation systems for fuel cells. *Power Technology*, 2021, 45(04): 545-551.
- [4] Zhangming Zhang. Design of Hydrogen Cycle Subsystem for High Power Fuel Cells. Tongji University, 2021.
- [5] Zhang L X, Li J, Li R Y, et al. A review of hydrogen supply system for automotive fuel cells. *Journal of Engineering Thermophysics*, 2022, 43(06): 1444-1459.
- [6] Tsai ShangWen, Chen YongSong. A mathematical model to study the energy efficiency of a proton exchange membrane fuel cell with a dead-ended anode. *Applied Energy*, 2017, 188: 151-159.

- [7] Bao Cheng, Ouyang Minggao, Yi Baolian. Modeling and control of air stream and hydrogen flow with recirculation in a PEM fuel cell system—I. Control-oriented modeling. *International Journal of Hydrogen Energy*, 2006, 31(13): 1879-1896.
- [8] Mohsen Dadvar, Ebrahim Afshari. Analysis of design parameters in anodic recirculation system based on ejector technology for PEM fuel cells: A new approach in designing. *International Journal of Hydrogen Energy*, 2014, 39(23): 12061-12073.
- [9] Ma Tiancai, Cong Ming, Meng Yixun, et al. Numerical studies on ejector in proton exchange membrane fuel cell system with anodic gas state parameters as design boundary. *International Journal of Hydrogen Energy*, 2021, 46(78): 38841-38853.
- [10] Shen YiWei. Characterisation of a rotary vortex hydrogen circulating pump. China University of Petroleum (Beijing), 2023.
- [11] He Lingxuan. Thermodynamic Analysis and Comprehensive Evaluation of Vehicle Fuel Cell Power System under Dynamic Operating Conditions. Hunan Institute of Science and Technology, 2023.
- [12] Wang Miao. Simulation and control strategy of hydrogen supply system for fuel cell engine. Shandong University, 2023. Barbir F. PEM Fuel Cells: Theory and Practice. Burlington: Elsevier/Academic Press, 2005.
- [13] Zhang DengHao. Influence of structural and operating parameters of a primer on the priming ratio. Donghua University, 2023.
- [14] Yinhai Zhu, Yanzhong Li. New theoretical model for convergent nozzle ejector in the proton exchange membrane fuel cell system. *Journal of Power Sources*, 2009, 191(2): 510-519.
- [15] Yang Z J. Deep learning-based dynamic modelling of automotive proton exchange membrane fuel cells. Tianjin University, 2022.

-30°C COLD START STRATEGY OF DESIGNED FUEL CELL SYSTEM

YinHao Yang*, JueXiao Chen, Chang Du

School of Automotive Studies, Tongji University, Shanghai 201804, China.

Corresponding Author: Yin hao Yang, Email: 3039822027@qq.com

Abstract: Proton exchange membrane fuel cells (PEMFCs) are considered one of the most promising alternative power sources for future vehicles due to their high energy conversion efficiency, zero pollution, and wide availability of fuel sources. Enhancing the low-temperature start-up capability of fuel cell systems is crucial for their widespread commercial application in the future. However, current experimental research findings are primarily based on single fuel cells or low-power stacks, with very limited studies on the impact of cold start on high-power systems. This leads to a significant gap between current scientific research and practical application, and the relevant results cannot be directly applied to actual systems. Therefore, research on low-temperature cold start of high-power fuel cells is of great significance. In this study, a 130kW fuel cell system was designed, and AVL Cruise M software was used to model and simulate the low-temperature cold start process of the fuel cell. By studying the start-up current loading strategies and the effects of operating parameters on fuel cell performance changes under -30°C experimental conditions, key information reflecting the state changes within the fuel cell stack was obtained. Based on this, a low-temperature cold start loading strategy corresponding to the specific temperature was proposed.

Keywords: PEMFC; -30°C cold start; Cruise M simulation; Loading strategy

1 INTRODUCTION

With the increasing severity of environmental pollution and energy crises, countries have successively taken measures to accelerate the development of clean and renewable energy sources, reduce the combustion of fossil fuels, and urgently seek new alternative clean energy sources. Proton exchange membrane fuel cells (PEMFCs), as one of the best applications of hydrogen energy, have attracted widespread attention. PEMFCs are characterized by their high energy density, environmental friendliness, noise-free operation, and clean efficiency. In recent years, significant breakthroughs have been made in the development of related materials and core components. However, the issue of cold start at low temperatures remains a major constraint on their commercialization and practical application.

PEMFCs initially contain liquid water, and liquid water is also generated during the cyclic operation process. When accumulated liquid water freezes in low-temperature environments ($<0^{\circ}\text{C}$), it can lead to several critical issues. The freezing of water can cover the gas diffusion layer with ice, preventing gases from reaching the surface of the catalyst layer and severely affecting gas transport. Ice formation in the catalyst layer can cover the reactive sites, reducing the electrochemical reaction active area. Additionally, freezing can cause localized stress on key materials and components within the fuel cell stack, damaging the cell's structural organization and causing permanent damage to the proton exchange membrane. Volume changes caused by freeze-thaw cycles can also lead to cracking in the catalyst layer, significantly affecting the performance of the fuel cell stack. These factors collectively result in a substantial decrease in the rate of electrochemical reactions, leading to cold start failure of PEMFCs at low temperatures and negatively impacting the lifespan and performance of the battery. Therefore, avoiding the extensive freezing of water during the low-temperature cold start process of PEMFCs is crucial for enhancing the performance of fuel cell systems.

Currently, both domestic and international research has been conducted on the cold start of fuel cells.

Regarding self-startup, various studies have explored different approaches to enhance the self-startup capabilities of fuel cells. In terms of constant-current self-startup, Zang compared the cold start performance of fuel cells under different current densities and proposed a boundary current density range for successful cold start at different temperatures [1]. In terms of variable-current self-startup, Gwak et al. achieved rapid cold start of the fuel cell by controlling the operating current during the cold start process [2]. Lei compared the effects of different current loading methods, such as constant current and linearly increasing current, on the cold start performance of fuel cells [3-4]. It was found that the cold start performance of fuel cells using variable current startup methods is significantly better than that of other startup methods.

Regarding assisted startup, in terms of coolant circulation heating, Ríos achieved successful cold start of the fuel cell at -30°C by heating the coolant [5]. Luo et al. studied the cold start issue of fuel cell vehicles with a coolant preheating strategy and realized successful cold start of the fuel cell at -30°C through coolant-assisted heating [6].

Regarding the loading strategy, Li et al. found that performing a large step load under a low current severely disrupts the uniformity of the cell voltage in the stack [7]. The impact of the loading magnitude on voltage uniformity is greater than that of the loading frequency. Migliardini et al. utilized a 6kW fuel cell stack to investigate the voltage uniformity under different constant loading and unloading rates [8].

However, current experimental research findings are primarily based on single fuel cells or low-power stacks. There is a significant lack of studies on the impact of cold start and the coupling of multiple parameters on cold start in high-

power systems. This has led to a considerable gap between the current scientific research and practical application, and the relevant results cannot be directly applied to the cold start performance evaluation and design of actual fuel cell systems. Therefore, this research will design a high-power fuel cell power system (130kW) and determine the effects of different loading methods and other parameter conditions on the cold start performance of fuel cells through experiments and simulations. Based on these findings, the startup strategy will be optimized.

2 DESIGN OF A FUEL CELL SYSTEM

Achieving low-temperature cold start of fuel cell systems is extremely important for their widespread application. In actual experimental testing, proton exchange membrane fuel cell (PEMFC) stacks exhibit a series of significant characteristics. First, PEMFCs have a large number of signals to be collected, requiring real-time monitoring and recording of numerous parameters to ensure stable system operation and performance evaluation. Second, control signals need to be fast and precise, as any slight delay or error can affect the performance and safety of the fuel cell. Additionally, the controlled components need to have rapid response capabilities to adapt to complex operating conditions and strong anti-interference capabilities to ensure stable operation in various complex environments.

During the development of fuel cell systems, the system design phase is crucial. It is necessary to ensure that the hardware and software performance can meet the high requirements of the fuel cell system and have good real-time capabilities to complete complex calculations and control tasks in a short time. This study employs a high-power fuel cell stack (comprising 418 single cells) to conduct low-temperature cold start experiments, and the relevant systems have been designed and constructed as shown in Figure 1, including the hydrogen supply system, the air supply system, the thermal management system, and the electrical and control system.

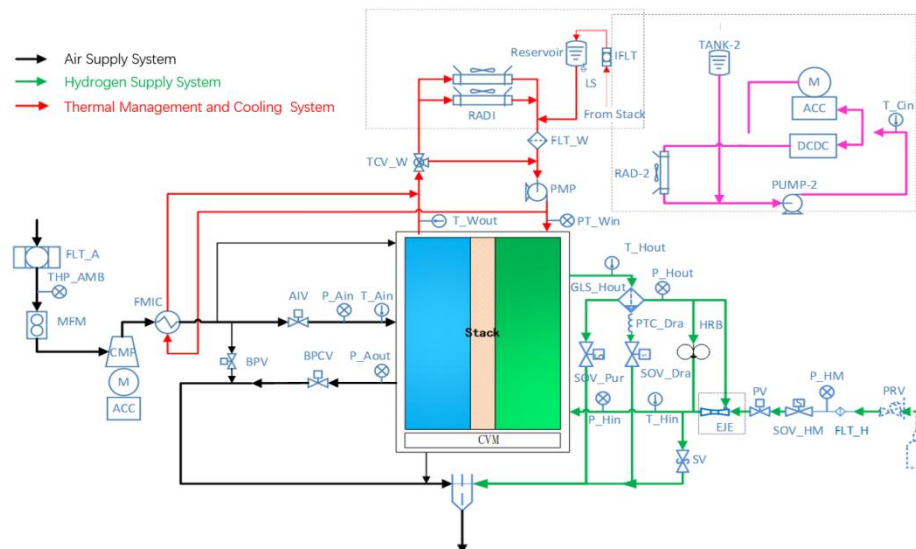


Figure 1 Schematic Diagram of the 130kW Fuel Cell System

2.1 The Air Supply System

The air supply system mainly consisting of an air filter, an integrated ambient temperature, humidity, and pressure sensor, an air mass flow meter, an air compressor, an intercooler, an air intake valve, and a backpressure valve. During normal operation of the fuel cell stack, air passes through the air filter and is compressed into high-temperature, high-pressure gas by the air compressor. It then enters the fuel cell stack after being cooled by the intercooler, providing the cathode of the stack with the oxidant.

The air filter selected is the Xuanke Hydrogen FC120 fuel cell cathode air filter. Xuanke Hydrogen's FC series fuel cell cathode air filters are safe, reliable, and available in a wide range of specifications. They are chemically adsorptive filter materials specifically developed for Chinese fuel cell systems, achieving an optimal balance between the filtration of harmful gases such as sulfur dioxide and nitrogen oxides, and dust particles. The product features high dust capacity, high adsorption efficiency, and simple, flexible installation methods.

The Xuanke Hydrogen FC120 fuel cell cathode air filter is suitable for fuel cell stacks ranging from 50 to 130 kW, with a rated flow rate of 720 slpm, which meets the requirements of this fuel cell stack. The detailed parameters are shown in the table. The air mass flow meter is used to monitor the air flow in the air path, thereby controlling the flow of the air compressor. The BOSCH 0281006270 model is selected, with an input voltage of 6–17 V and a rated flow rate of 640 kg/h (with a measuring range of -60 to 800 kg/h). The pressure drop is 12 kPa, and it is equipped with an NTC.

The integrated ambient temperature, humidity, and pressure sensor is used to detect the temperature, humidity, and pressure of the air at the air intake of the fuel cell stack. The Wuxi Shengbang ST8251-1BBA1 type temperature, humidity, and pressure integrated sensor is selected, and its product characteristics are shown in Table 1. The air

compressor is a crucial component of the air supply system, providing the necessary air pressure and flow rate for the fuel cell stack. The Haidowell HEC30 air compressor is selected for this purpose. It features a maximum motor power of 30 kW, a maximum flow rate of 180 g/s, a maximum pressure ratio of 3.3, and a maximum rotational speed of 120,000 revolutions per minute.

Table 1 Product Characteristics of Integrated Sensor

Characteristic	Range
Temperature Measurement Range	-40°C~125°C
Relative Humidity Output	0%RH~100%RH
Pressure Measurement Range	20kPa(A)~300kPa(A)

2.2 The Hydrogen Supply System

The hydrogen supply system includes components such as hydrogen storage cylinders, pressure-reducing valves, medium-pressure solenoid valves, hydrogen recirculation pumps, drain valves, and proportional valves. The hydrogen exhaust valve selected is the ASCO X256548494 with a 2.0 mm orifice. This valve has a Kv value of 0.129, the maximum hydrogen exhaust flow rate is 233.3 slpm.

It is estimated that the anode chamber volume of a 418-cell stack is approximately 1.672 L. From this, it can be inferred that opening this hydrogen exhaust valve for 0.43 seconds each time would be sufficient to replace the gas inside the chamber once, meeting the usage requirements. The drain valve selected is the ASCO X986542315 with a 3.5 mm orifice.

The medium-pressure solenoid valve is defined as the switch between the first-stage pressure reduction of hydrogen and the proportional valve. The model selected is the Zhejiang Hongsheng HONGSHGN-FSV-1. The role of the proportional valve is to provide the hydrogen pressure and flow rate required for the normal operation of the fuel cell stack. The proportional valve chosen is the Weifu WHI22.

The purpose of the hydrogen recirculation pump is to improve the utilization rate of hydrogen and increase the hydrogen flow rate, thereby enhancing the water drainage efficiency.

The medium-pressure sensor is used to detect the pressure after the pressure-reducing valve. The model selected is the Sensata 32CP42-01-ENV, with a pressure range of 0.1-2.0 MPa(A) and an operating temperature of -40°C to +125°C. The low-pressure sensor is used to detect the hydrogen pressure from the proportional valve outlet to the fuel cell stack inlet. The model selected is the Sensata 30CP42-06-ENV, with an operating pressure of 50–300 kPa(A). The rated hydrogen inlet pressure of the fuel cell stack is 266.30 kPa(A), which meets the usage requirements. The hydrogen inlet and outlet temperature sensors are used to detect the temperature of the hydrogen at the inlet and outlet of the fuel cell stack. The model selected is the Qufu Tianbo Fuel Cell Temperature Sensor 1927, with an operating temperature range of -40°C to 140°C. The water separator is used to separate liquid water from the wet hydrogen at the anode to prevent liquid water from entering the anode of the fuel cell stack and causing anode flooding. The model selected is the Suzhou Ruidu HWS120-WC gas-water separator. The safety valve is used to release pressure when the pressure after the proportional valve is too high, thus protecting the fuel cell stack. The safety valve is a purely mechanical device. Based on experience, the set pressure = the rated hydrogen absolute pressure of the fuel cell stack \times 1.1. Calculations show that the inlet pressure of the fuel cell stack should not exceed 2.926 bar (2.66×1.1). The Zhejiang Hongsheng HSXYF-121L meets this requirement. The hydrogen recirculation pump selected is the Suzhou Ruidu WDE-C008-H hydrogen recirculation pump, with its specific performance parameters shown in Table 2.

Table 2 Product Characteristics of HRP

Characteristic	Range/Value
Maximum Suction Pressure	230kPa(A)
Maximum Discharge Pressure	300kPa(A)
Pressure Ratio Range	1-1.2
Displacement	200CC
Maximum Volumetric Flow Rate	750L/min
Speed Range	500-700rpm

2.3 The Thermal Management System

The electrochemical reactions occurring inside the fuel cell stack generate heat. Excessive heat can cause the stack temperature to rise too high, leading to performance degradation of the stack components. In severe cases, it can reduce the service life of the stack. The primary function of the thermal management system is to maintain the thermal balance

of the fuel cell system, dissipate the excess heat generated by the stack, ensure that the stack quickly reaches a suitable temperature, and prevent the stack from overheating.

The thermal management system mainly consists of a coolant pump, an expansion tank, a thermostat, and a radiator (with a cooling fan). The thermostat controls the large and small coolant circulation loops. When the stack temperature is low, the thermostat directs the coolant to flow within the small circulation loop. When the coolant temperature is high, the thermostat directs the coolant to flow through the radiator assembly for cooling, thereby maintaining the normal operating temperature of the stack.

The coolant pump controls the coolant flow rate by adjusting its speed to achieve temperature control, ensuring that the fuel cell system operates within an appropriate temperature range. The cooling fan transfers the heat from the coolant to the environment, reducing the coolant temperature. The cooling fan is required to have a high airflow rate, low noise, stepless speed control, and the ability to feedback its operating status.

The circulating water pump selected is the Beijing Ai'er LQY-P150 model pump. The ion filter is connected in series between the fuel cell stack coolant exhaust port and the expansion tank branch, and the I2M i10-3 ion filter is chosen. The temperature and pressure integrated sensor is used to detect the coolant temperature and pressure at the fuel cell stack inlet. Based on the detected data, the circulating water pump and fan are adjusted in speed to ensure that the fuel cell operates at the recommended working temperature and to maintain the coolant pressure within the normal range. The temperature and pressure integrated sensor selected is the Sensata 31CP02-03-ENV.

2.4 The Electrical and Control System

The function of the electrical and control system is to ensure that other systems can operate efficiently and in coordination, guaranteeing sufficient gas supply and appropriate working temperature, among other things. It is mainly composed of various types of sensors, flow meters, valve components, and so on.

3 COLD START SIMULATION AT -30°C

3.1 Model Assumption

To facilitate the calculations, the following assumptions are made for the PEMFC model to simplify the analysis within the stack:

- 1) All gases within the model are assumed to be ideal gases, and the influence of gravity is neglected.
- 2) The initial state of water generated by the electrochemical reaction is assumed to exist in the cathode catalyst layer in the form of membrane-bound water.
- 3) Only diffusion, heat transfer, and mass transfer in the direction perpendicular to the plane are considered.
- 4) Pressure variations within the fuel cell are ignored.
- 5) Liquid water and ice in the flow channels are neglected.
- 6) The heat exchange between the fuel cell stack and the external environment is assumed to be uniform, i.e., the phenomenon where the heat dissipation rate at the ends of the stack is much higher than other parts, leading to lower temperatures of the end cells in the stack, is ignored.

3.2 Model Construction

The relevant parameters of the fuel cell stack are shown in Table 3. Based on the software AVL Cruise M, a cold start simulation model that reflects the internal state of the PEMFC is established, and its structure is shown in Figure 2.

Table 3 The Parameters of The Fuel Cell Stack

Parameter	Value
Number of Cell	418
Reaction Active Area	330cm ²
Thickness of Membrane	8μm
Thickness of Bipolar Plate	2mm
Flow Channel Cross-section	1mm ²
Flow Channel Length	500mm

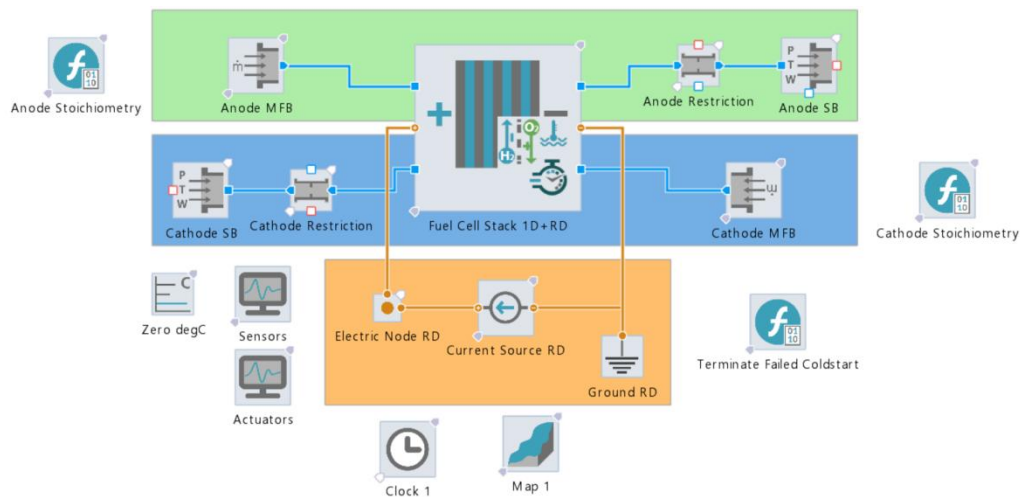


Figure 2 The Model of the 130kW Fuel Cell System

3.3 Simulation Result

Figure 3 illustrates the current loading process in the simulation, which is designed to closely match the experimental loading curve.

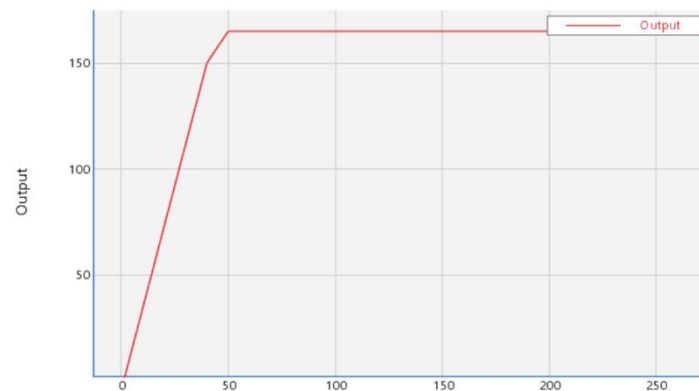


Figure 3 Current Loading in Simulation

The cell voltage variation after simulation is shown in Figure 4. Since the simulation is conducted on the cell stack model, while the experiment is performed on the entire system, there is inevitably some deviation. Additionally, the voltage fluctuations in the simulation are relatively large. Therefore, only a qualitative analysis can be made based on this simulation. It can be observed that the battery voltage slightly drops after startup and then stabilizes after a period of fluctuation. The fact that the voltage does not drop to 0V indicates that the cold start is successful.

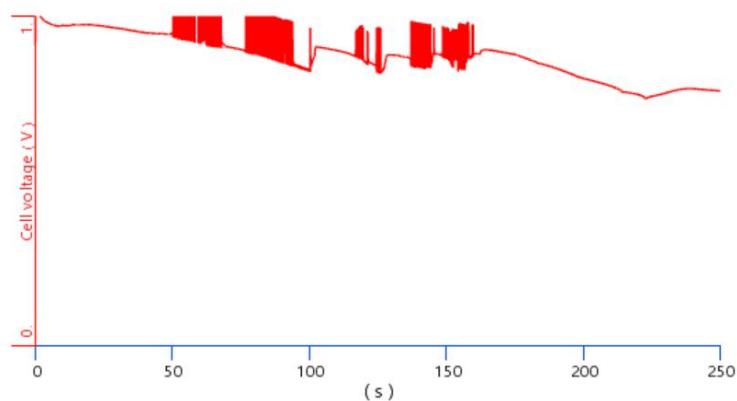


Figure 4 Cell Voltage during Simulation

4 COLD START EXPERIMENT AT -30°C

4.1 Experimental Procedure

The cold start experimental procedure is as follows:

- 1) Set the environmental chamber to a specific temperature and place the fuel cell system inside for more than 12 hours.
- 2) After the temperature stabilization is complete, turn on the electronic load and open the hydrogen inlet valve. Check for any hydrogen leaks at the valve.
- 3) Once all preparations are complete, send a cold start command to the fuel cell system. The control unit will perform a self-check on the system, verifying the communication status of system components such as the air compressor, hydrogen recirculation pump, and water pump to ensure they are operating normally. Then, the cold start program will be executed, and the system will begin loading. During the warm-up process, as the coolant temperature increases, the system power will gradually rise to the set power level.
- 4) After the system has operated stably at the set power for a period of time, the shutdown procedure will be executed. Subsequently, the stack load will decrease to idle load, and the purging program will start to expel any residual water.
- 5) Once the purging is complete, turn off the electronic load and stabilize the system at the set temperature for another 12 hours until the next cold start experiment begins.

4.2 Experimental Result

The Voltage and Current Curve, Coolant Temperature Curve, Voltage Range Curve during startup at -30°C is shown in Figure 5. As can be seen from Figure 5, the use of a continuous loading strategy allows the system to successfully cold start at -30°C. The current is loaded from 0A to approximately 150A within 40 seconds at a rate of 3.75A/s. During this period, the temperature of the stack's coolant outlet rises steadily and exceeds 0°C, with a temperature increase rate of 0.73°C/s. During the loading process, the average voltage of the stack and the minimum voltage of the single cell decrease. Once the coolant temperature exceeds 0°C, the voltage begins to recover and continues to rise until the system operates stably. At this point, the difference between the average voltage and the minimum single cell voltage is negligible.

It can also be observed from the range chart that, at the initial stage of loading, the range begins to increase gradually, reaching a maximum of approximately 0.38V. After the coolant temperature exceeds 0°C, the range starts to decrease. When the system operates stably, the range is essentially zero, indicating good consistency of the stack and good system output performance.

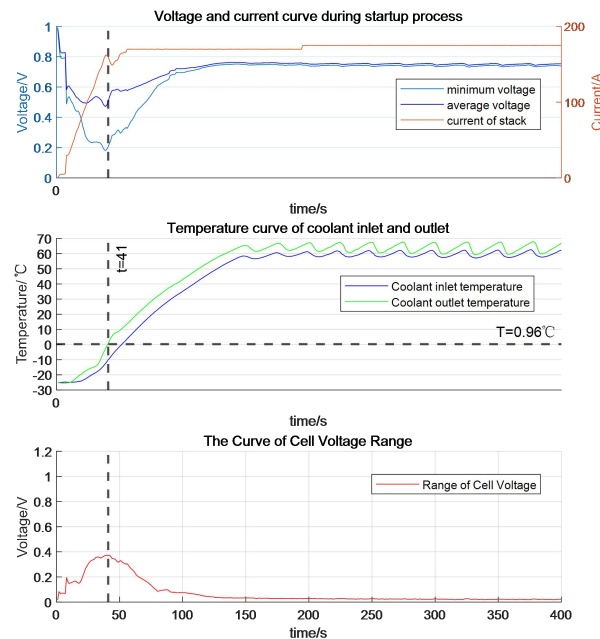


Figure 5 The Voltage and Current Curve, Coolant Temperature Curve, Voltage Range Curve during startup at -30°C

The voltage distribution of individual cells at a current of 175A is shown in Figure 6. It can be seen that in the initial stage of current loading, the voltage of the lowest single cell drops rapidly, but it does not fall below the protection threshold. During the fast continuous loading period, the coolant temperature rises at a relatively high rate. About 41

seconds after loading, the coolant outlet temperature of the cell stack exceeds 0°C. The drop in the lowest voltage and average voltage continues until this time. After that, the coolant temperature continues to rise, while the lowest voltage and average voltage begin to recover. When the coolant temperature rises to about 50°C, the lowest voltage and average voltage return to a stable level, with little difference between the two.

It can also be seen from the voltage difference graph of individual cells that before the coolant temperature rises to 0°C, the voltage difference gradually increases, but it does not exceed 0.4V. After the coolant temperature rises to 0°C, the voltage difference gradually decreases until it reaches zero. Both the voltage difference graph and the individual cell voltage graph show good voltage consistency of the cell stack.

The impedance variation during the shutdown and purge process after operation is shown in Figure 7. The purge process lasts for about 200 seconds, with the highest purge impedance reaching 1082 mΩ.

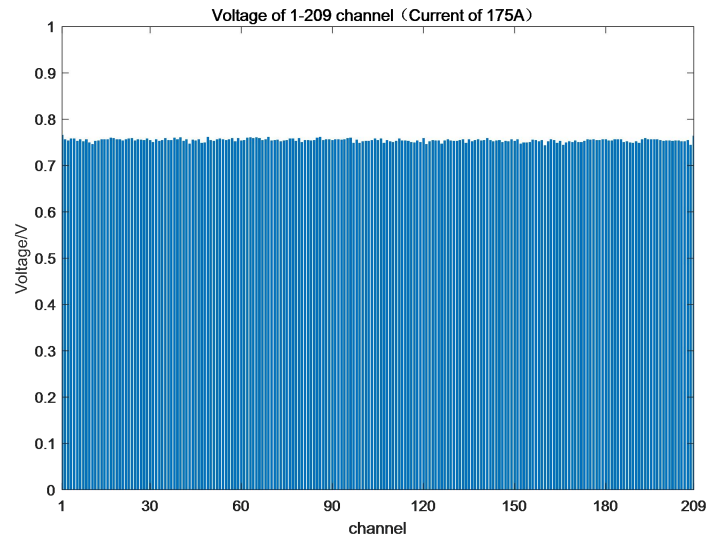


Figure 6 The Voltage and Current Curve, Coolant Temperature Curve, Voltage Range Curve during startup at -30°C

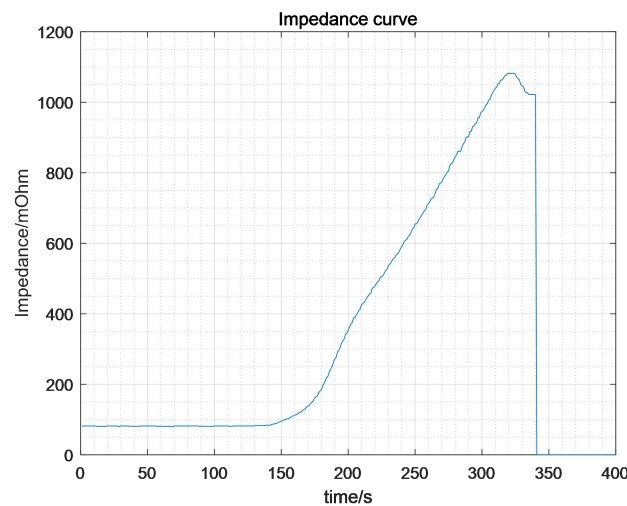


Figure 7 The Voltage and Current Curve, Coolant Temperature Curve, Voltage Range Curve during startup at -30°C

5 CONCLUSION

This paper designs a fuel cell integrated system for low-temperature cold start and verifies the correctness and effectiveness of the startup strategy used in this paper through simulation and experiments. The results show that rapid continuous loading at a rate of 3.75A/s is conducive to heat generation in the stack and the increase in coolant temperature, and enables successful cold start at -30°C. Of course, there are also shortcomings in this paper. For example, the subsequent simulation analysis can be more quantitative rather than limited to qualitative analysis.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

FUNDING

This work is supported by the National Natural Science Foundation of China (Grant No.52077157).

REFERENCES

- [1] Zang L, Hao L. Numerical study of the cold-start process of PEM fuel cells with different current density operating modes. *Journal of Energy Engineering*, 2020, 146(6): 04020057.
- [2] Gwak G, Ju H. A rapid start-up strategy for polymer electrolyte fuel cells at subzero temperatures based on control of the operating current density. *international journal of hydrogen energy*, 2015, 40(35): 11989-11997.
- [3] Lei L, He P, He P, et al. A comparative study: The effect of current loading modes on the cold start-up process of PEMFC stack. *Energy Conversion and Management*, 2022, 251: 114991.
- [4] Lei L, He P, He P, et al. Numerical Research on the Cold Start-up Strategy of a PEMFC Stack from -30° C. *Journal of Thermal Science*, 2022: 1-13.
- [5] Ríos G M, Schirmer J, Gentner C, et al. Efficient thermal management strategies for cold starts of a proton exchange membrane fuel cell system. *Applied Energy*, 2020, 279: 115813.
- [6] Luo M, Zhang J, Zhang C, et al. Cold start investigation of fuel cell vehicles with coolant preheating strategy. *Applied Thermal Engineering*, 2022, 201: 117816.
- [7] Li M, Dai C H, Guo A, et al. Experimental study on dynamic voltage uniformity of a 2 kW aircooled PEMFC. *Electrical Engineering*, 2018, 100: 2725-35. 79.
- [8] Migliardini F, Palma T D, Gaele M, et al. Cell voltage analysis of a 6 kW polymeric electrolyte fuel cell stack designed for hybrid power systems. *Materials Today: Proceedings*, 2019, 10(3): 393- 399.

CONSTRUCTION OF CLASSIFICATION METHOD FOR URBAN ROAD INTERSECTIONS

Lai Wei, XiChan Zhu, ZhiXiong Ma*

School of Automotive Studies, Tongji University, Shanghai, 201804, China.

Corresponding Author: ZhiXiong Ma, Email: mzx1978@tongji.edu.cn

Abstract: In order to scientifically and reasonably test and evaluate the driving ability and behavior of intelligent vehicles in urban scenarios, it is necessary to first cover all types of scenes for the most complex intersection scenes in the city. However, due to the current research mostly using simple intersection shapes for scene classification, it is difficult to achieve exhaustive and traversal of all types of scenes at intersections. This article innovatively proposes the use of road rights composed of a limited number of combinations of directional arrows and traffic signals in the lane as the classification basis for intersections. Through research and summarization, all combination types are obtained, and the frequency distribution of all combination types is statistically analyzed in the automobile demonstration area. Based on this, the classification is carried out, laying the foundation for the testing and evaluation of intelligent vehicles in urban working conditions.

Keywords: Testing and evaluation; Intersection classification; Urban scene; Right of way

1 INTRODUCTION

Unlike highway traffic scenes, urban road traffic scenes are mainly divided into road section driving and intersection traffic. Among them, intersections are very complex traffic scenes, with not only complex road structures but also numerous traffic participants and uncontrollable factors, making them prone to traffic accidents. Therefore, vehicles not only need to predict intersections and record driving routes, but more importantly, understand the intersection scene and choose appropriate driving strategies to prevent typical traffic incidents from occurring. Therefore, for the testing and evaluation of urban scenarios, if you want to construct typical evaluation scenarios at intersections. It is necessary to first use effective classification methods to classify the intersection scene.

The most common classification method currently is to classify scenes based on the shape of intersections. Zhou Jianhua et al.[1] simply divided intersections into cross shaped, circular, X-shaped, T-shaped, Y-shaped and other forms. Liu Chunxu[2] proposed to divide typical intersections into cross shaped intersections, T-shaped intersections, roundabout intersections, Y-shaped intersections, misaligned intersections, X-shaped intersections, and multi way intersections. At the same time, complex intersections are divided into distorted Y-shaped complex intersections, distorted roundabout complex intersections, distorted cross shaped complex intersections, distorted 5-way complex intersections, and distorted 4-way complex intersections. Ying Shen et al.[3] provided a classification of intersections, which are classified into the following categories based on geometric shapes: Y-shaped intersections, T-shaped intersections, cross intersections, X-shaped intersections, roundabouts, and compound intersections; According to the structure, the intersection is divided into non channelized intersection and channelized intersection. Later, it was proposed to decompose complex intersection scenes into multiple sub scenes for autonomous vehicle driving decisions, select corresponding sub scenes based on driving intentions, match intersection classifications, and help cars better understand intersection scenes.

Ma Xuehan et al.[4] proposed a method for classifying intersections based on natural driving research, which summarizes the elements of intersection traffic into geometric type, traffic control type, and lane type. For convenience, each type is represented by a code, that is, a combination of letters can be used to determine a certain intersection type. This method is relatively scientific, but the classification is too simple to cover all roads in the city. In addition, he also analyzed several intersection traffic scenarios through China FOT video statistics and obtained a distribution table of intersection types.

In terms of intersection classification, the Ministry of Transport of the People's Republic of China has released multiple standards that provide detailed explanations and classifications of intersection geometry types, traffic control types, and lane types. Road Traffic Signs and Markings Part 2: Road Traffic Signs "[5] classifies directional signs, including straight ahead, left turn, right turn, straight ahead and left turn, straight ahead and right turn, left and right turn, roundabout driving, and other directional signs. The basic shapes and meanings of common directional arrows are given in "Road Traffic Signs and Markings Part 3: Road Traffic Markings"[6], and some arrows are shown in Figure 5. The "Specification for the Setting and Installation of Road Traffic Signal Lights"[7] lists the combination forms of motor vehicle signal lights and direction indicator signals, and divides them into conventional and special situations. Partial situations are shown in Figure 6 and Figure 7. The above relevant standards of the Ministry of Transport provide ideas and references for subsequent scene classification.

In summary, current research mostly uses simple intersection shapes for scene classification. However, different numbers of lanes and signal lights within the same shape at intersections can lead to vastly different scenes, making it difficult to exhaustively and traverse all types of scenes using intersection shape classification. Even the subsequent

proposal to decompose complex intersection scenes into multiple sub scenes still adopts the intersection shape classification method, which is consistent with the above problem. Later, researchers proposed a more scientific and effective method of classifying intersections based on their geometric types, traffic control types, and lane types. However, at that time, the research was relatively general and lacked refinement and comprehensiveness. This article takes this as a breakthrough point, seeking to use a limited number of indicators to classify intersections, striving to achieve an exhaustive list of urban intersection scene types.

2 CLASSIFICATION IDEAS FOR URBAN ROAD INTERSECTIONS

As mentioned earlier, this article breaks away from the previous focus on the structure of intersections, the number of lanes at intersections, the number of directions at intersections, and selects a limited number of representative and universal relevant indicators as the basic basis for classification. By observing urban road intersections, it can be observed that although there may be differences in the number of lanes and directions at the intersection, the basic driving behaviors performed by a single vehicle passing through the intersection are limited, with a maximum of four driving behaviors: straight ahead, left turn, right turn, and U-turn.

The execution of each driving behavior at the intersection is determined by the road rights formed by the direction indicated by the directional arrows in front of the intersection stop line and the clear priority of traffic signals at the intersection. The types of directional arrows and traffic signals in the lane are limited and can be exhaustively listed. Therefore, no matter what kind of intersection can be encountered on urban roads, the road rights formed by the combination of directional arrows and traffic signals in the lane can be used as the classification basis for intersections.

Therefore, it is necessary to first count all types of directional arrows and traffic lights that have appeared in the city, including conventional and special types. Then, all types of directional arrows and traffic lights can be combined in sequence to cover all possible road rights at intersections, achieving full coverage of urban road intersections and laying the foundation for the construction of subsequent intelligent driving intersection evaluation schemes.

3 CLASSIFICATION COLLECTION RESULTS OF URBAN ROAD INTERSECTIONS

Due to the use of dual indicators of directional arrows and traffic signals as classification criteria for intersections in urban roads, it is necessary to first exhaustively list the types of these two indicators at the intersection, and then combine the collected results of the two indicators to classify the intersection.

3.1 Types of Directional Arrows at Urban Road Intersections

This section mainly summarizes the types of directional arrows that exist on conventional motor vehicle lanes at intersections on urban roads, excluding non motor vehicle lanes and lanes for unconventional vehicles such as bus lanes. At the intersection of urban roads, due to the provisions of "Road Traffic Signs and Markings Part 3: Road Traffic Markings" (GB 5768.3-2009), there are 9 basic shapes and meanings of directional arrows for intersections, including multiple bidirectional directional arrows in addition to single directional arrows, covering most common types of standards.

In addition, the regulations also include two special types of lanes: variable lanes and tidal lanes. Variable lanes are generally marked with text within the lane, and indicator signs or signal lights are set up above or in front of the variable lane to indicate the direction in which the current lane can be driven. The tidal lane, with a double yellow dashed line composed of two parallel yellow dashed lines as its indicator line, generally uses corresponding variable signs and lane direction signal control facilities to cooperate and achieve the function of indicating the current lane direction.

However, in some complex road sections, such as intersections with complex geometric shapes or limited number of lanes, standard directional arrows may not be able to fully convey information. At this point, the traffic management department will make judgments based on actual needs and engineering, so there may be some non-standard directional arrows at intersections on real city roads to better guide drivers. Therefore, when calculating the types of directional arrows at intersections, it is also necessary to summarize unconventional types that are not included in regulations. Through extensive field research and online searches, except for situations where there are no directional arrows at intersections, the following eight non-standard types were extracted, including seven types of directional arrows and one type of lane:

- (1) No U-turn: Indicate that no U-turn is allowed ahead (as shown in Figure 1)
- (2) No Left Turn: Indicates that no left turn is allowed ahead (as shown in Figure 2)
- (3) No Right Turn: Indicates that right turns are prohibited ahead (as shown in Figure 3)
- (4) Straight (limited time no left turn): Indicates that vehicles are prohibited from turning left in front of the lane for a specific period of time and can only proceed straight (as shown in Figure 4)
- (5) Left turn and no U-turn: Indicates that only left turns are allowed and no U-turns are allowed ahead (as shown in Figure 5)
- (6) Straight ahead+left turn+right turn: indicates that you can go straight ahead or turn left or right (as shown in Figure 6)
- (7) Left turn, straight ahead, no U-turn: indicates that you can go straight ahead or turn left, but not make a U-turn (as shown in Figure 7)

(8) Borrowing left turn lane: Left turning vehicles borrow the opposite lane to make a left turn, usually using supporting facilities such as traffic signals, LED screens, signs and markings to indicate the right of way of the current lane (as shown in Figure 8).



Figure 1 No U-turn



Figure 2 No Left Turn



Figure 3 No Right Turn



Figure 4 Straight (Limited Time No Left Turn)



Figure 5 Left Turn and No U-turn



Figure 6 Straight Ahead+Left Turn+Right Turn

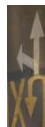


Figure 7 Left Turn, Straight Ahead, No U-turn








Figure 8 Borrowing Left Turn Lane

Based on the above conventional and unconventional types, the number of directions that can be controlled simultaneously according to the directional arrows can be divided into zero level, first level, second level, third level, and special type. For convenience, each type of directional arrow is represented in uppercase code form, with most codes using the first letter of the English alphabet as a shorthand, such as Straight, Left, Right, U-turn, and multi-level categories mostly in the form of overlapping first letters. The specific categories, names, codes, diagrams, and directional arrows of the classification are shown in Table 1.

Table 1 Classification of Directional Arrows at Urban Road Intersections

Category	Name	Code	Sketch map
----------	------	------	------------







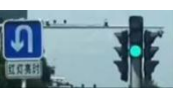









Level 0	Unmarked line	N	
	Straight	S	
	Left	L	
	Right	R	
Level 1	U-turn	U	
	No U-turn	nU	
	No Left Turn	nL	
	No Right Turn	nR	
Level 2	Left turn+U-turn	LU	
	Straight+Left turn	SL	
	Straight+Right turn	SR	
	Straight+U-turn	SU	
	Left turn+Right turn	LR	
	Left turn and No U-turn	LnU	
	Straight (limited time no left turn)	SpnL	
	Straight ahead+Left turn+Right turn	SLR	
Level 3	Left turn, straight ahead, no U-turn	SLnU	
	Changed lane	CL	
Special type	Tide lane	TL	
	Borrowing left turn lane	JL	

3.2 Types of Motor Vehicle Signal Lights at Urban Road Intersections

Similarly, through statistics and induction, a combination classification table of motor vehicle signal lights at urban road intersections can be obtained, as shown in Table 2.

Table 2 Classification Table of Motor Vehicle Signal Light Combinations at Urban Road Intersections

Category	Name	Code	Sketch map
----------	------	------	------------

Standard	Single circular signal light	d	
	Circular+left turn signal light	ld	
	Circular+right turn Signal Light	rd	
	Circular+left turn+right turn signal light	ld	
	Straight+left turn+right turn signal light	lsr	
	No/constant yellow signal light	n	
Non-standard	U-turn specific signal light	u	
	Circular+U-Turn signal light	ud	
	Circular+left turn+U-turn signal light	uld	
	Straight+right turn+U-turn signal light	usr	
	Lane signal light	cd	
	Circular+left turn&U-turn signal light	lu_d	
Special type	Straight+left turn+left turn&U-turn signal light	lu_ls	
	Changed lane signal light	cl	
	Tide lane signal light	tl	
	Borrowing left turn lane signal light	jl	

3.3 Summary of Types of Urban Road Intersections

After obtaining 20 types of directional arrows and 16 types of motor vehicle signal lights at the intersection mentioned above, the combination of the two can be used to determine the right of way at the intersection and effectively classify the intersection. The classification results after combination are shown in Table 3. The code naming and right of way for each combination type are provided in Table 3. The blue code represents the type of road rights that are not controlled by traffic lights, while the red code represents the type of road rights that are not scientific or reasonable enough. Therefore, the types with red codes are temporarily not included in the common intersection types.

According to statistics, there are a total of 189 combination codes, of which 82 are red codes. Therefore, there are 107 common types of intersections, including 23 blue code types that are not controlled by traffic lights. Therefore, when analyzing intersections in the future, the types of intersections summarized above can be used for analysis.

Table 3 Classification Table of Urban Road Intersection Types

Directional arrows	Signal light	Standard								Non-standard				Special type			
	Code	d	ld	rd	ldr	lsr	n	u	ud	uld	usr	cd	lu_d	lu_ls	cl	tl	jl
Level 0	N	d-N	ld-N	rd-N	ldr-N	lsr-N	n-N	/	ud-N	uld-N	usr-N		lu_d-N	lu_ls-N			
	S	d-S	ld-S	rd-S	ldr-S	lsr-S	n-S	/	ud-S	uld-S	usr-S		lu_d-S	lu_ls-S			
	L	d-L	ld-L	rd-L	ldr-L	lsr-L	n-L	/	ud-L	uld-L	/		lu_d-L	lu_ls-L			
Level 1	R	d-R	ld-R	rd-R	ldr-R	lsr-R	n-R	/	ud-R	uld-R	usr-R		lu_d-R	lu_ls-R			
	U	d-U	ld-U	rd-U	ldr-U	lsr-U	n-U	u-U	ud-U	uld-U	usr-U		lu_d-U	lu_ls-U			
	nU	d-nU	ld-nU	rd-nU	ldr-nU	lsr-nU	n-nU	/	ud-nU	uld-nU	usr-nU		lu_d-nU	lu_ls-nU			
	nL	d-nL	ld-nL	rd-nL	ldr-nL	lsr-nL	n-nL	/	ud-nL	uld-nL	usr-nL		lu_d-nL	lu_ls-nL			
	nR	d-nR	ld-nR	rd-nR	ldr-nR	lsr-nR	n-nR	/	ud-nR	uld-nR	usr-nR		lu_d-nR	lu_ls-nR			
	LU	d-LU	ld-LU	rd-LU	ldr-LU	lsr-LU	n-LU	/	ud-LU	uld-LU	usr-LU	/	lu_d-LU	lu_ls-LU			
	SL	d-SL	ld-SL	rd-SL	ldr-SL	lsr-SL	n-SL	/	ud-SL	uld-SL	usr-SL		lu_d-SL	lu_ls-SL			
Level 2	SR	d-SR	ld-SR	rd-SR	ldr-SR	lsr-SR	n-SR	/	ud-SR	uld-SR	usr-SR		lu_d-SR	lu_ls-SR			
	SU	d-SU	ld-SU	rd-SU	ldr-SU	lsr-SU	n-SU	/	ud-SU	uld-SU	usr-SU		lu_d-SU	lu_ls-SU			
	LR	d-LR	ld-LR	rd-LR	ldr-LR	lsr-LR	n-LR	/	ud-LR	uld-LR	usr-LR		lu_d-LR	lu_ls-LR			
	LnU	d-LnU	ld-LnU	rd-LnU	ldr-LnU	lsr-LnU	n-LnU	/	ud-LnU	uld-LnU	/		lu_d-LnU	lu_ls-LnU			
Level 3	SpnL	d-SpnL	ld-SpnL	rd-SpnL	ldr-SpnL	lsr-SpnL	n-SpnL	/	ud-SpnL	uld-SpnL	usr-SpnL		lu_d-SpnL	lu_ls-SpnL			
	SLR	d-SLR	ld-SLR	rd-SLR	ldr-SLR	lsr-SLR	n-SLR	/	ud-SLR	uld-SLR	usr-SLR		lu_d-SLR	lu_ls-SLR			
	SLnU	d-SLnU	ld-SLnU	rd-SLnU	ldr-SLnU	lsr-SLnU	n-SLnU	/	ud-SLnU	uld-SLnU	usr-SLnU		lu_d-SLnU	lu_ls-SLnU			
Special type	CL														cl_CL		
	TL															tl_TL	
	JL																jl_JL

4 CLASSIFICATION AND PROPORTION OF URBAN ROAD INTERSECTIONS

After obtaining the above classification types of intersections, research can be conducted on each type, and the frequency and distribution density of each type in the urban road environment can be calculated. This can further obtain the urban coverage of each intersection type and divide them according to frequency.

In the intersection research stage, the first step is to select a suitable research scope. Due to the need to fully consider the diversity of the number and distribution of intersections within the scope and the operability of the research work, combined with the representativeness of the selected area, it is decided to choose Shanghai Intelligent Connected Vehicle Demonstration Zone in China as the research object. On the one hand, the scope of the demonstration zone is relatively limited, which can achieve exhaustive and research work on all intersections within the range. On the other hand, as a venue support for the development of intelligent connected vehicle technology in China, the intelligent connected vehicle demonstration zone is determined to promote the research and application of intelligent connected vehicle technology. It covers numerous testing scenarios and application scenarios on test roads, and can fully serve as a representative area of urban road environment.

Combining online statistics and field research to complement each other, this article summarizes a total of 1863 combinations of 297 smart intersections in the Shanghai Automobile City Experimental Zone. According to the previous naming rules, the number and frequency of each combination type are all counted. The statistical results are shown in Table 4.

Table 4 Types of Urban Intersection Combinations in Shanghai Automobile City Experimental Zone

Type	Count	Frequency
d-N	311	16.69%
ld-S	228	12.24%
d-SR	215	11.54%
d-S	160	8.59%
ld-L	161	8.64%
n-N	156	8.37%
d-SL	146	7.84%
d-L	147	7.89%
ld-SR	124	6.66%
d-R	74	3.97%
ld-R	24	1.29%
n-S	25	1.34%
n-SR	16	0.86%
n-L	9	0.48%

d-LR	9	0.48%
ld-LnU	7	0.38%
ldr-S	6	0.32%
n-SL	7	0.38%
d-nU	6	0.32%
n-R	4	0.21%
ldr-L	4	0.21%
rd-R	3	0.16%
ldr-SR	3	0.16%
rd-S	2	0.11%
ldr-R	2	0.11%
d-LnU	2	0.11%
ld-LR	2	0.11%
n-LR	2	0.11%
rd-L	2	0.11%
ldr-U	1	0.05%
ld-LU	1	0.05%
d-SLnU	1	0.05%
d-SLR	1	0.05%
ld-SL	1	0.05%
d-T1	1	0.05%
Total	1863	

By analyzing the frequency of each type, it was found that three types, d-N, ld-S, and d-SR, had a frequency greater than 10%, with a total of 9 types exceeding 5% and 12 types exceeding 1%. When categorizing intersection combination types based on frequency, combination types greater than 10% are defined as extremely common types, 5% -10% are defined as relatively common types, 1% -5% are defined as common types, 0.1% -1% are defined as uncommon types, and less than 0.1% are defined as rare types.

5 CLASSIFICATION EXTRACTION RESULTS OF URBAN ROAD INTERSECTIONS

The purpose of classifying urban road intersections is to achieve full coverage of all types of intersections on urban roads with limited classification methods and quantities, providing a basis for developing evaluation methods for urban navigation assistance systems. In the testing and evaluation process, different evaluation criteria need to be developed for the difficulty level of driving at different intersections. Therefore, it is necessary to classify all intersection combinations into levels based on their frequency of occurrence and difficulty level.

Firstly, referring to the classification of test scenarios under high-speed road conditions, in the 2020 revised version of the Navigation Intelligent Driving Evaluation Regulations[8], the test scenarios are divided into four types: basic scenarios, challenge scenarios, innovation scenarios, and backup scenarios. At urban road intersections, this article divides the collected combinations of urban intersections in the experimental zone into three testing scenarios: basic scenarios, challenge scenarios, and innovative scenarios. Among them, scenes that occur frequently and are relatively simple for existing vehicles will be classified as basic scenes; Classify scenes with moderate frequency of occurrence and certain challenges and difficulties as challenge scenes; Classify low frequency, challenging, and difficult scenarios as innovative scenarios.

Based on the frequency of occurrence and driving difficulty of the 34 reasonable existence types excluding d-T1 in the Shanghai Automobile City test zone, the test scenarios are divided as shown in Table 5. Among them, 23 types are basic scenarios, 8 are challenge scenarios, and 3 are innovation scenarios.

Table 5 Classification of Test Scenarios for Intersection Types in Shanghai Automotive City Experimental Zone

	d-N	ld-S	d-SR	d-S	ld-L	n-N	d-SL	d-L
Basic scenario	ld-SR	d-R	n-S	n-SR	n-L	ldr-S	n-SL	n-R
	ldr-L	rd-R	ldr-SR	rd-S	ldr-R	n-LR	rd-L	
Challenge scenario	ld-R	d-LR	ld-LnU	d-nU	d-LnU	ld-LR	ld-LU	d-SLR
Innovative scenario	ldr-U	d-SLnU	ld-SL					

6 CONCLUSION

In response to the problem of full coverage of scene types for the most complex intersection scenes in cities, this article innovatively proposes using the road rights composed of a limited number of combinations of guide arrows and traffic lights in the lanes as the classification basis for intersections. Through research and summarization, all types of combinations of guide arrows and traffic lights at intersections are first obtained, and then all combination types are obtained by pairwise combination. The frequency distribution of all combination types is analyzed through statistics of the automobile demonstration zone, and based on this, the classification is carried out, laying the foundation for the testing and evaluation of intelligent vehicles in urban work conditions.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

- [1] Zhou Jianhua, Song Rui, Liu Mingyang, et al. Research on Key Technologies of Autonomous Driving Perception and Recognition in Intersection Scenarios. *China Automotive*, 2022 (05): 32-38.
- [2] Liu Chunxu. Research on Real time Signal Intelligent Control Technology for Complex Intersections. Chongqing Jiaotong University, 2013.
- [3] Ying Shen, Liang Yuanyi, Jiang Yuewen, et al. Intersection classification and vehicle driving methods and equipment adapted to autonomous driving. Hubei Province: CN115366887A, 2022.
- [4] Ma Xuehan, Zhu Xichan, Ma Zhixiong. Analysis of traffic classification and violations in natural driving research//State Key Laboratory of Advanced Design and Manufacturing for Vehicle Body, Hunan University. Infots Procedures of the 14th International Forum of Automotive Traffic Safety, 2017: 28-41
- [5] GB 5768.2-2022, Road traffic signs and markings - Part 2: Road traffic signs. 2022. Retrieved from <https://openstd.samr.gov.cn/bzgk/gb/newGbInfo?hcno=15B1FC09EE1AE92F1A9EC97BA3C9E451>
- [6] GB 5768.3-2009, Road traffic signs and markings Part 3: Road traffic markings. 2009. Retrieved from <https://std.samr.gov.cn/gb/search/gbDetailed?id=71F772D7C970D3A7E05397BE0A0AB82A>
- [7] GB 14886-2016, Specification for Setting and Installation of Road Traffic Signal Lights. 2016. Retrieved from <https://std.samr.gov.cn/gb/search/gbDetailed?id=71F772D8163DD3A7E05397BE0A0AB82A>
- [8] IVISTA National Intelligent Automotive Integration System Test Zone IVISTA China Intelligent Vehicle Index Navigation Intelligent Driving Evaluation Regulations (2020 Revised Edition) [S/OL]. 2020. Retrieved from <https://www.i-vista.org/d/file/p/2022-09-27/98f534dad920e6f4515376d98b21d69f.pdf>

STREETSCAPE GREENNESS AND PARK SERVICE EVALUATION IN ZHENGZHOU, CHINA: A SPATIAL MULTI-ZONING PERSPECTIVE

Da Mao^{1*}, ZhiYu Yuan¹, MengLei Zhao¹, HuaYu Fu²

¹*School of Horticulture and Landscape Architecture, Henan Institute of Science and Technology, Xinxiang 453003, Henan, China.*

²*Xinxiang Santian Landscape Engineering Co., Ltd., Xinxiang 453003, Henan, China.*

Corresponding Author: Da Mao, Email: maoda@foxmail.com

Abstract: This study investigates the spatial association between streetscape greenness (Green View Index, GVI) and park service evaluations in Zhengzhou, China, integrating multi-source geospatial data, including 131 park POIs and 23,866 street view images with a multi-zoning analytical framework (grid-based, radial-sector, and Voronoi zoning). Using spatial autocorrelation (Global/Local Moran's I) and bivariate LISA cluster analyses, key findings include: (1) Radial-sector zoning outperformed other methods in capturing spatial heterogeneity; (2) High GVI clusters concentrated in the urban core, while top-rated parks exhibited concentric patterns, with low-value zones coupled at the periphery; (3) Significant synergy emerged between peak park ratings and mean GVI ($I = 0.135\text{--}0.196$), revealing asymmetric interactions. A three-tiered planning strategy is proposed: radial-sector green space allocation, targeted upgrades in mismatch zones, and flagship park-driven green networks. This research advances methodological innovation in green infrastructure optimization for urbanizing cities.

Keywords: Streetscape; Green view index; Multi-zoning; Park service

1 INTRODUCTION

Urban green spaces, encompassing streetscapes and parks, serve as critical infrastructure for advancing ecological sustainability, promoting public health outcomes, and addressing urban social equity challenges. Despite their shared objectives in urban greening, parks and streetscapes exhibit distinct spatial service units that rarely overlap geographically. This spatial divergence has historically hindered comprehensive investigations into their service correlations, particularly under technological constraints of spatial multi-source data scarcity in earlier decades [1, 2]. The emergence of geospatial big data and computational advancements now enables precise quantification of public service facility distributions through innovative metrics such as street view image analytics and Point of Interest (POI) data mining, revolutionizing urban landscape assessments [3].

The proliferation of web-mapping services (e.g., Google Street View, Baidu Map, Amap, Tencent Map) has democratized access to streetscape imagery, facilitating large-scale urban analyses at unprecedented granularity [4]. Seminal work by Li et al. pioneered the use of Google Street View (GSV) imagery with fisheye projection algorithms to quantify pedestrian-scale green exposure [5]. This methodology laid the foundation for the Green View Index (GVI), a standardized metric quantifying visible vegetation proportions in streetscapes through semantic segmentation of street-level imagery [6-8]. Concurrently, POI data has emerged as a vital tool for urban analytics, enabling researchers to: Map facility distribution patterns, Assess urban functional vitality hotspots, and Evaluate service catchment areas of green spaces through spatial interaction modeling [9-12].

The methodological core of this study lies in addressing the spatial mismatch between streetscape greenness service radii and park accessibility through a multi-zoning analytical framework. We operationalize this approach by:

Calculating neighborhood-level GVI scores using street view imagery, Quantifying park service performance through POI-based visitation patterns and facility quality metrics, and applying spatial regression models to identify zone-specific associations between streetscape greenness and park service efficacy. This study employs a multi-zoning analytical framework to test diverse spatial unit delineation methods. The underlying scientific hypothesis posits that the Green View Index (GVI)—quantifying streetscape greenness—exhibits statistically significant correlations with public evaluations of park services, with such associations being spatially scale-dependent. Specifically, we postulate that the strength and direction of GVI-park service relationships vary across different hierarchical spatial units (e.g., 1000m grids, voronoi area, concentric zone area).

2 DATA AND METHODS

2.1 Study Area

This study focuses on Zhengzhou City, the provincial capital of Henan Province, strategically positioned as a key socioeconomic and transportation hub in Central China. As a rapidly urbanizing megacity, Zhengzhou has witnessed over 50% expansion in built-up areas during the past two decades, while persistent disparities in green space

accessibility and quality remain unresolved. These dynamics position Zhengzhou as a critical case for examining the challenges of reconciling green infrastructure development with population growth (exceeding 10 million residents) and spatial intensification.

The study area encompasses the urban core within Zhengzhou's Third Ring Road (Figure 1), comprising four expressway segments: the North Third Ring Road (N3RR), East Third Ring Road (E3RR), South Third Ring Road (S3RR), and West Third Ring Road (W3RR). This enclosed 202 km² zone represents the city's primary urbanized territory, characterized by systematic street networks, multi-tiered park systems, and representative streetscape greenery configurations.

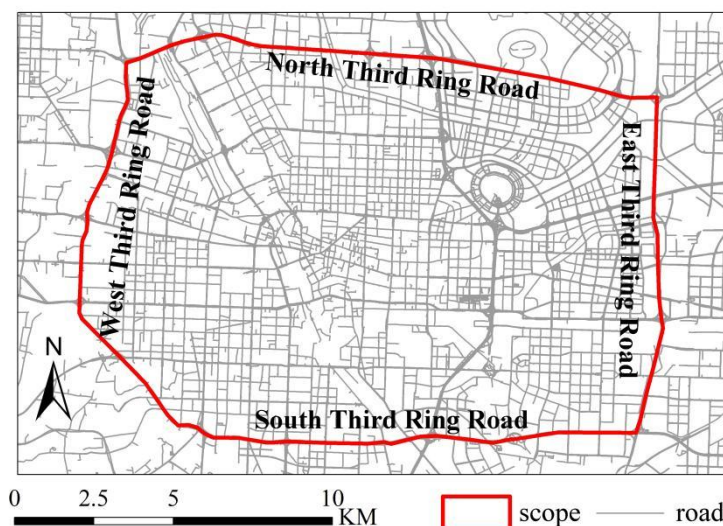


Figure 1 Scope Diagram of Study Area

Edited on this map source: <https://www.openstreetmap.org/>

2.2 Data

2.2.1 POI data and park rating score

The POI data in this study were sourced from AutoNavi Map (Amap), China's leading digital mapping platform, with all records timestamped for 2023 to ensure temporal consistency with concurrent streetscape imagery. Amap was prioritized over alternative providers due to its unique integration of crowdsourced facility ratings—a critical feature enabling quantitative assessment of public perception toward park services. Each park's composite rating, ranging from 0 (lowest) to 5 (highest), was calculated as the arithmetic mean of individual user evaluations, reflecting collective satisfaction levels.

Through protocol-driven web scraping techniques, we systematically collected georeferenced POI records for 131 municipal parks within Zhengzhou's Third Ring Road boundary. The automated harvesting pipeline included:

- (1) Spatial filtering: Restricting queries to parks located within the 202 km² urban core.
- (2) Data validation: Eliminating duplicate entries to ensure statistical reliability.
- (3) Attribute extraction: Capturing essential metadata (park name, GPS coordinates, rating score).

2.2.2 Street view images and GVI

The streetscape imagery was sourced from Baidu Map, China's leading provider of panoramic street view services. It is critical to note that while data acquisition occurred between December 2022 and January 2023, the actual capture dates of Baidu Map images may span multiple years due to the platform's asynchronous update cycles. A systematic sampling protocol yielded 23,866 georeferenced street view points within the study area, spaced at 50-meter intervals along road networks. The technical process (Figure 2) is as follows:

Step 1: Programmatic Street View Data Harvesting

Georeferenced 360° panoramic images were systematically retrieved through Baidu Map's API (v4.0) using a spatially stratified sampling strategy. To comply with platform rate limits, we implemented token-bucket throttling algorithms (50 requests/second) while ensuring complete coverage across all road segments within Zhengzhou's Third Ring Road. Each panorama was acquired at 4,096×2,048 resolution (RGB, 8-bit depth).

Step 2: Cubemap Projection via PTGui Professional

The acquired 23,866 street view panoramas underwent systematic geometric processing following established protocols in urban visual perception studies [13-15]. Using PTGui Pro's batch processing module, each spherical panorama was decomposed into six perspective-corrected cube faces (front/back/left/right/top/bottom) through equiangular cubemap projection, which generated 143,196 directional images (23,866×6). Outputs were standardized as 960×960 JPG tiles per face (8-bit RGB).

Step 3: Human-Centric Visual Field Simulation

To ensure ecological validity in simulating pedestrian visual experiences, zenith and nadir projections (47,732 images) were excluded based on empirical evidence that vertical extremes contribute minimally to human horizontal field-of-view perception [16]. The final dataset comprised 95,464 street view images ($23,866 \times 4$). For pedestrian visual emulation, the upper 2/3 portion (960×720 pixels) was extracted from front/back/left/right faces. This cropping strategy aligns with:

- (1) Eye-level perspective: 1.2-1.8 m vertical focus matching human height percentiles.
- (2) Urban context preservation: Maximizes street-level vegetation visibility.

Step 4: GVI Computation via Deep Learning Pipeline

A pretrained PSPNet-101 model (initialized with ADE20K weights) was fine-tuned on our UrbanGreen-15K dataset (containing 12 vegetation classes) using transfer learning. The processing pipeline included:

- (1) Semantic segmentation: Pixel-wise classification at 512×512 resolution
- (2) Vegetation masking: Thresholding greenness (HSV ranges: $H = 80-160^\circ$, $S > 20\%$, $V > 15\%$)
- (3) GVI calculation:

$$GVI = \left(\frac{\text{Vegetation Pixels}}{\text{Total Valid Pixels}} \right) \times 100\% \quad (1)$$

Ultimately, the GVI of each point is the average of the GVI of the four directions' images.

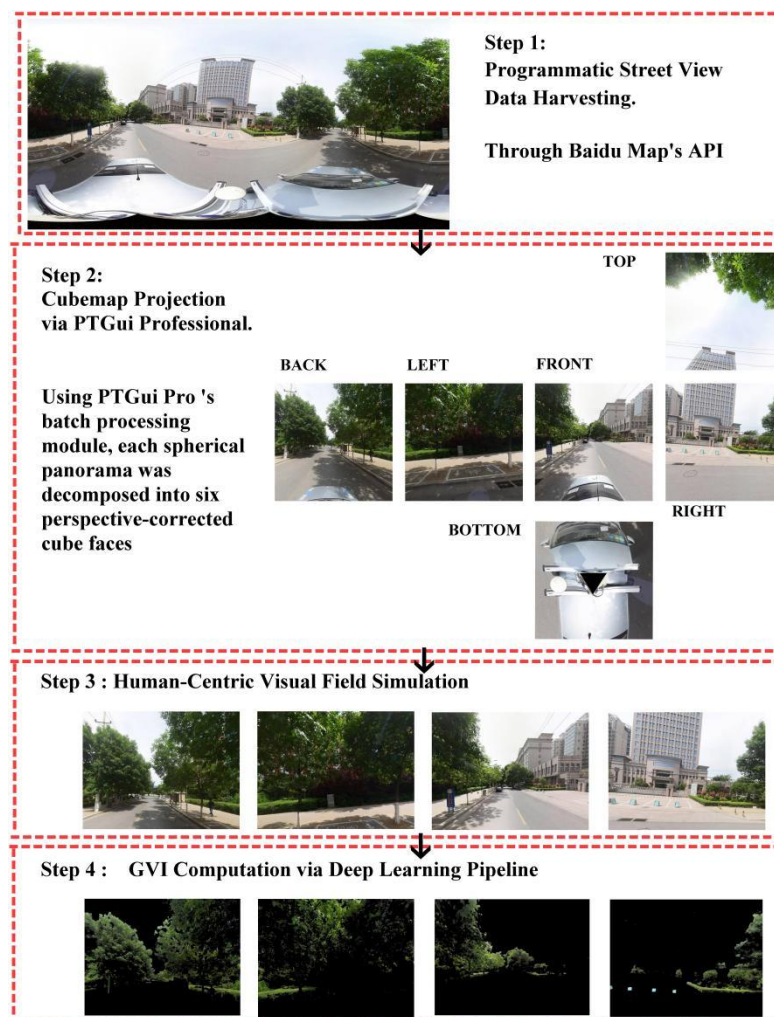


Figure 2 Technical Flow chart

2.3 Methods

2.3.1 Spatial multi-zoning

The multi-zoning framework constitutes the methodological nucleus of this investigation. We systematically implemented and compared three distinct spatial zoning paradigms (Figure 3) to address scale-dependent heterogeneity in green service associations:

(1) Grid-based Zoning

Adopting a regular square tessellation approach, the study area was partitioned into $1 \text{ km} \times 1 \text{ km}$ grid cells (1 km^2 each), aligned with the Universal Transverse Mercator (UTM) Zone 50N coordinate system. This isotropic partitioning facilitates density-based spatial analysis while minimizing shape-induced biases, particularly suitable for examining city-wide greenness distribution patterns.

(2) Radial-sector Zoning

Centered on Zhengzhou's urban green heart (Zijinshan Park), we constructed concentric buffer rings at 1 km radial intervals extending outward. Each annular zone was further subdivided into eight cardinal and intercardinal sectors (N, NE, E, SE, S, SW, W, NW), creating combined radial-sector units. This dual-axis partitioning enables directional analysis of green space diffusion patterns and distance-decay modeling of park service effectiveness.

(3) Voronoi Zoning

The Voronoi diagram (also termed Thiessen polygons), named after mathematician Georgy Voronoi, was employed to delineate park service boundaries through spatial partitioning. Geometrically, Voronoi diagrams are constructed by generating perpendicular bisectors between adjacent control points, forming continuous polygonal cells where any location within a cell is closer to its generating point than to others [16]. Operationalizing this theoretical framework, we generated Voronoi polygons using 131 park POIs from Amap within Zhengzhou's Third Ring Road through ArcGIS' Voronoi tool. This approach defines the maximum theoretical service extent for each park, assuming users prioritize nearest-facility access.

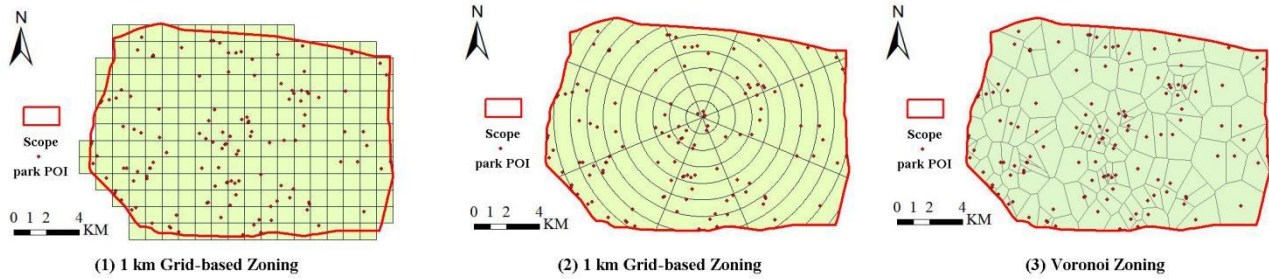


Figure 3 Three Distinct Spatial Zoning Paradigms

2.3.2 Spatial unit attribute assignment

This study incorporated two primary datasets for spatial unit characterization: park ratings derived from POI evaluations, and streetscape greenness measured through Green View Index (GVI).

(1) GVI Integration:

With dense sampling points (23,866 locations at 50m intervals), all spatial units contained sufficient GVI measurements. Using ArcGIS Pro's spatial join tool, three greenness metrics were computed for each unit: MeanGVI, MaxGVI and MinGVI.

(2) Park Rating Interpolation:

Given sparse POI distribution (131 parks), non-Voronoi zoning methods (grid/radial) contained null units. To address this spatial discontinuity, we implemented ordinary Kriging interpolation (spherical semivariogram model, 0.5 km search radius) to create continuous rating surfaces. The interpolated raster was subsequently integrated with spatial units via zonal statistics, yielding Unit Rating.

2.3.3 Spatial correlation methodology

The values of the above spatial units were analyzed by applying the method of spatial correlation. Exploratory Spatial Data Analysis (ESDA) provides a suite of statistical techniques to quantify and visualize spatial autocorrelation patterns. This study employs both global and local indicators to examine the spatial association between streetscape greenness and park service evaluations.

(1) Global Spatial Autocorrelation

The *Moran's I* statistic (Moran, 1950) measures overall spatial dependence across the study area:

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (2)$$

where:

n : Number of spatial units

w_{ij} : Spatial weight matrix (queen contiguity)

S_0 : Sum of all spatial weights

x_i : Attribute value at location

\bar{x} : Mean attribute value

Interpretation:

$I \in [-1, 1]$: Positive values indicate clustering (similar values aggregate), negative values dispersion, and 0 random distribution. Statistical significance tested via permutation tests (999 simulations, $\alpha=0.05$).

(2) Local Indicators of Spatial Association (LISA)

Developed by Anselin (1995), LISA identifies localized clustering patterns and spatial outliers through: *Local Moran's I* computation for each spatial unit:

$$I_i = \frac{(x_i - \bar{x})}{S^2} \sum_{j=1}^n w_{ij} (x_j - \bar{x}) \quad (3)$$

where S^2 is the variance of x .

3 RESULTS

3.1 Univariate Global Spatial Autocorrelation Analysis

The Global *Moran's I* statistic was computed using GeoDa 1.20 to quantify spatial clustering patterns of streetscape greenness (GVI) under three distinct zoning frameworks:

Table 1 The Univariate Global *Moran's I*

Zoning Method	MeanGVI	MaxGVI	MeanGVI	Mean Park rating score	Max Park rating score	Min Park rating score
Grid	0.407	0.469	0.214	0.351	0.698	-0.121
Radial-sector	0.405	0.362	0.047	0.401	0.314	0.534
Voronoi	0.346	0.399	0.162	0.286	0.716	0.400

As can be seen from the univariate global *Moran's I* table (Table 1) above, the performance of several variables differs across the three spatial zoning methods. Among them, Mean GVI, Max GVI, Mean Park Rating Score, and Max Park Rating Score exhibit significant positive spatial autocorrelation in all three zoning frameworks. However, Min GVI and Min Park Rating Score show insignificant spatial clustering. Numerically, the Radial-sector method demonstrates the best overall performance, indicating that Radial-sector is a highly suitable approach.

3.2 Bivariate Local *Moran's I*

Using GeoDa 1.20's bivariate Local *Moran's I* (LISA) module, we systematically analyzed spatial cross-correlations between paired variables (e.g., GVI vs. park ratings). Representative cases are described below.

(1) 1 km Grid-based Zoning

Within the 1 km grid-based zoning spatial units, no significant bivariate spatial autocorrelation was observed between Mean Park Rating Score and Mean GVI. However, relatively significant spatial autocorrelation features were identified between:

Mean Park Rating Score and Max GVI (*Moran's I* = 0.153).

Max Park Rating Score and Mean GVI (*Moran's I* = 0.135).

In the LISA map (Figure 4) of Mean Park Rating Score vs. Max GVI, High-High (H-H) clusters exhibited distinct concentric patterns, while GVI itself demonstrated a spatial gradient characterized by higher values in the urban core and lower values in peripheral areas. Similarly, the LISA map (Figure 5) of Max Park Rating Score vs. Mean GVI revealed annular distributions of H-H and Low-High (L-H) units. These patterns indicate that while streetscape greenness (GVI) peaks in Zhengzhou's central urban area, park service ratings display annular differentiation features.

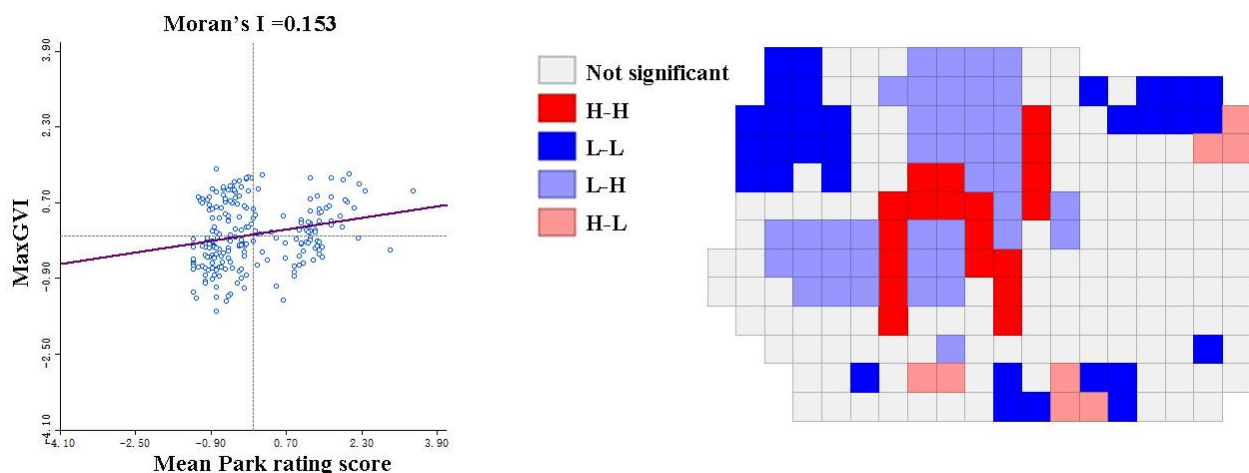


Figure 4 Moran Scatter Plot (left) & LISA Cluster Map of Mean Park Rating Score and MaxGVI (Right)

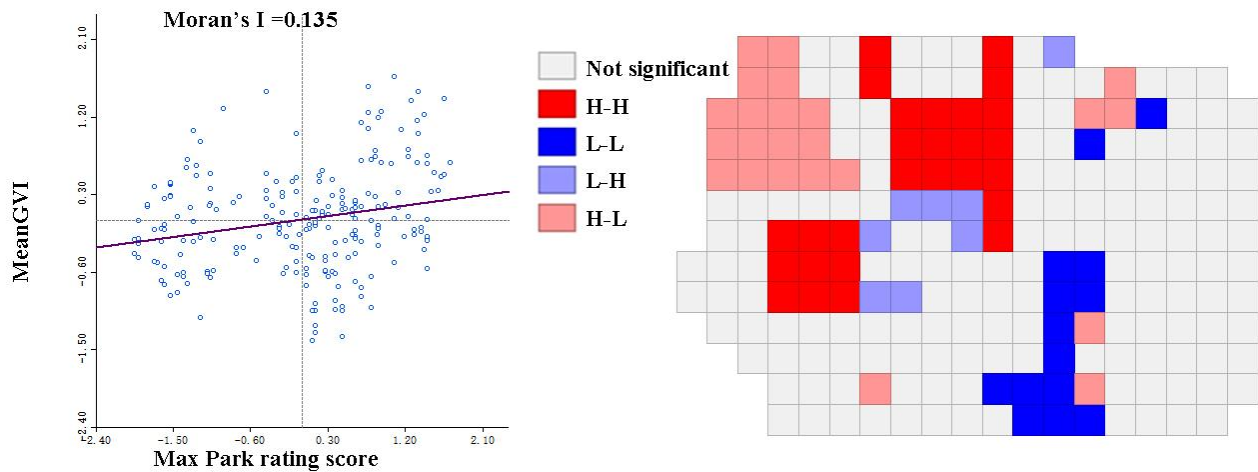


Figure 5 Moran Scatter Plot (left) & LISA Cluster Map of Max Park Rating Score and Mean GVI (Right)

(2) 1 km Radial-sector Zoning

Within the 1 km Radial-sector Zoning spatial units, relatively pronounced bivariate spatial autocorrelation was observed between Mean Park Rating Score and Mean GVI (*Moran's I* = 0.169). Similarly, Max Park Rating Score and Mean GVI also exhibited significant spatial autocorrelation (*Moran's I* = 0.216).

In the LISA map (Figure 6) of Mean Park Rating Score vs. Mean GVI, High-High (H-H) clusters were concentrated in the central radial sectors, while Low-Low (L-L) units predominantly aggregated in the east-central sectors. The distribution of H-H and L-L units in the LISA map (Figure 7) of Max Park Rating Score vs. Mean GVI mirrored this pattern. These findings demonstrate that the 1 km Radial-sector Zoning method effectively captures the concentric spatial stratification of urban units.

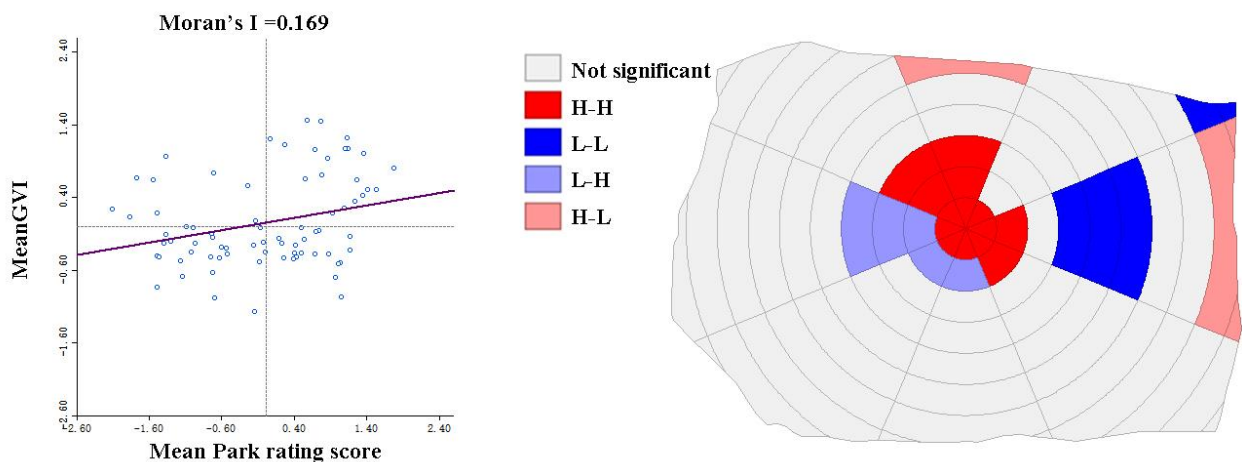


Figure 6 Moran Scatter Plot (left) & LISA Cluster Map of Mean Park Rating Score and Mean GVI (right)

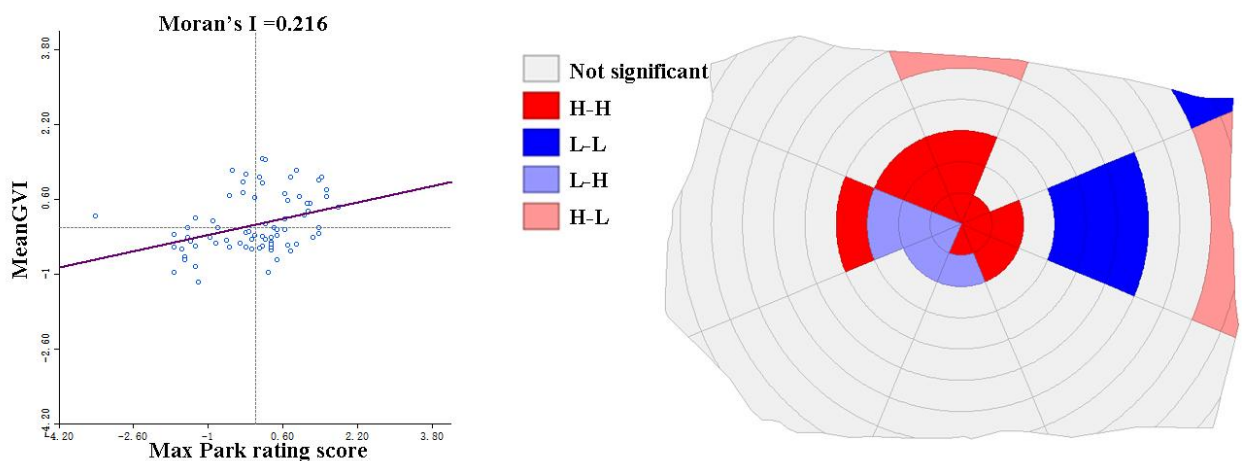
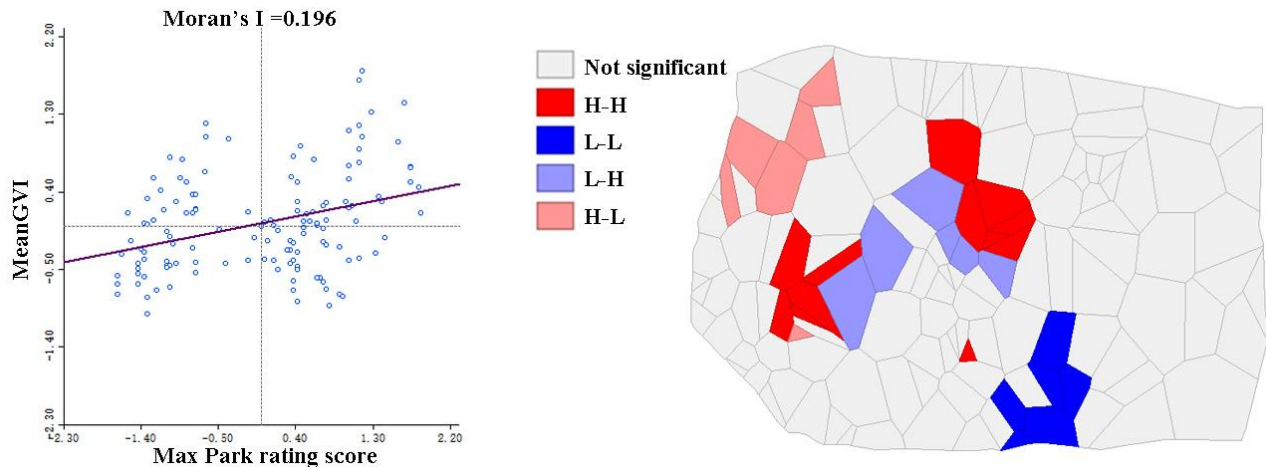
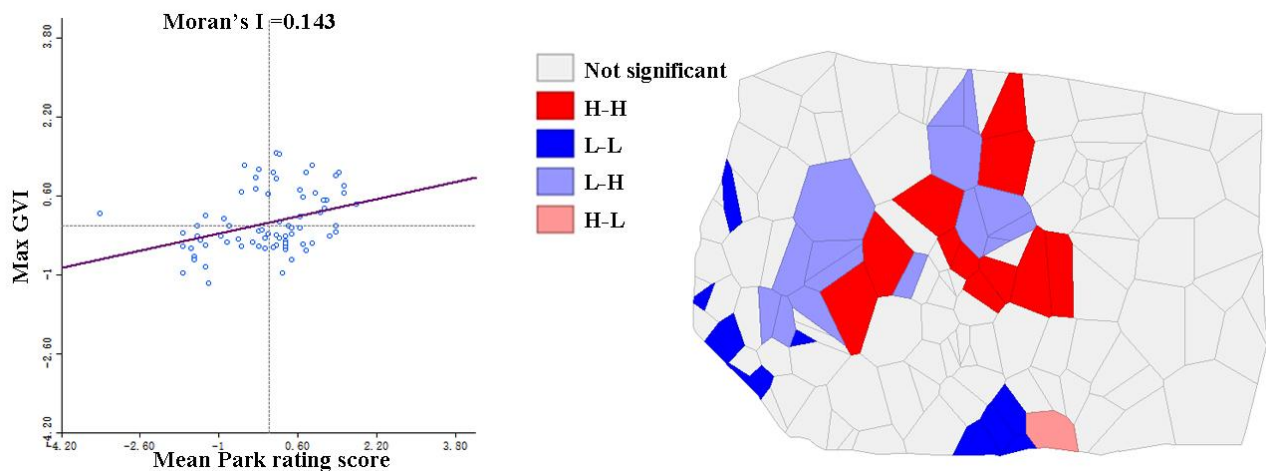


Figure 7 Moran Scatter Plot (left) & LISA cluster map of Max Park rating score and Mean GVI (right)**(3) Voronoi Zoning**

Within the Voronoi Zoning spatial units, no significant bivariate spatial autocorrelation was detected between Mean Park Rating Score and Mean GVI ($Moran's I = 0.033$). However, Max Park Rating Score and Mean GVI exhibited relatively significant spatial autocorrelation ($Moran's I = 0.196$). Notably, the spatial morphology of units in Voronoi diagrams is challenging to characterize systematically (Figure 8). In the LISA map (Figure 9) of Mean Park Rating Score vs. Max GVI, High-High (H-H) and Low-High (L-H) clusters were concentrated in central areas, while Low-Low (L-L) and High-Low (H-L) units were scattered across the southwestern periphery. These results suggest that while Voronoi Zoning can analyze spatial correlations, its morphological interpretation is less intuitive compared to the previous two zoning methods.

**Figure 8** Moran Scatter Plot (left) & LISA Cluster Map of Max Park Rating Score and Mean GVI (Right)**Figure 9** Moran Scatter Plot (left) & LISA Cluster Map of Mean Park Rating Score and Max GVI (Right)

4 CONCLUSIONS

The Zhengzhou case study yielded the following key findings:

(1) Spatial Association between Park Services and Streetscape Greenery

A significant correlation exists between park service evaluations and streetscape greenness. Crowdsourced park ratings (derived from POI data) and Green View Index (GVI) metrics (extracted from street view imagery) effectively captured this relationship, with $Moran's I$ values ranging from 0.135 to 0.216 ($p < 0.05$) under optimal zoning frameworks.

(2) Comparative Efficacy of Zoning Methods

All three spatial partitioning approaches — Grid-based, Radial-sector, and Voronoi Zoning — demonstrated utility in analyzing park-greenery associations. However, Radial-sector Zoning exhibited superior analytical capacity, achieving the highest $Moran's I$ values for both univariate and bivariate spatial autocorrelation analyses.

(3) Spatial Gradient Patterns

Streetscape Greenery: High-GVI clusters formed contiguous agglomerations in the urban core, reflecting centralized green infrastructure investments. **Park Services:** High-rating parks exhibited concentric distributions, aligning with

Zhengzhou's radial-concentric development model. Low-Value Zones: Peripheral areas showed synchronized deficiencies in both GVI and park ratings, highlighting urban-rural green equity gaps.

(4) Asymmetric Interaction Mechanisms

While no significant correlation emerged between mean park ratings and mean GVI, robust associations were observed between Peak park ratings and mean GVI, Mean park ratings and peak GVI. This asymmetry suggests that superior streetscape greenness serves as a foundational element for high park service ratings, while flagship parks can elevate neighborhood perceptions of green quality.

5 DISCUSSIONS

This study underscores the critical role of spatial zoning framework selection in urban green space analysis. While Voronoi partitioning demonstrated mathematical-theoretical rigor, it faced interpretive limitations in capturing gradient patterns. These findings advocate for context-sensitive zoning protocols, suggesting future exploration of hierarchical hexagonal spatial indexing and temporally weighted Voronoi models incorporating POI dynamics.

Practically, we propose a three-tiered spatial strategy for Zhengzhou:

- (1) Strategic development of wedge-shaped green corridors aligned with anisotropic urban expansion along primary development axes to enhance sectoral spatial configurations, urban ventilation, and ecological connectivity.
- (2) Precision interventions in peripheral mismatch zones—intensifying street greening in Low-High (L-H) clusters (GVI: 0.18 – 0.25, Park Ratings: 2.1 – 3.4) and modernizing recreational facilities in High-Low (H-L) areas.
- (3) Strategic anchoring through flagship parks, establishing 500-meter radiating greenway networks with pedestrian/bicycle priority design standards, coupled with community co-management initiatives.

This integrated approach synthesizes analytical rigor with implementable planning solutions, addressing both scale-dependent spatial heterogeneity and green service equity through adaptive governance mechanisms and multi-stakeholder participatory frameworks.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

FUNDING

This research was funded by the following projects: Cultivation Plan of Young Backbone Teachers in Colleges and Universities of Henan Province, China (2021GGJS118), Key Science and Technology Research and Development Program of Henan Province, China (232102320180), Key Science and Technology Research and Development Program of Henan Province, China (222102320233).

REFERENCES

- [1] Tang Z L, Gu S. An Evaluation of Social Performance in the Distribution of Urban Parks in the Central City of Shanghai: From Spatial to Social Equity. *Urban Planning Forum*, 2015(2): 48-56.
- [2] Huang Z, Tang L, Qiao P, et al. Socioecological justice in urban street greenery based on green view index- A case study within the Fuzhou Third Ring Road. *Urban Forestry & Urban Greening*, 2024. DOI: 10.1016/j.ufug.2024.128313
- [3] Wang H, Zhang W, Song H, et al. Spatial Evolution of Urban Population in Changsha and Its Simulation: Based on the Multi-source Data. *Economic Geography*, 2023, 43(8): 49-61.
- [4] Cheng L, Chu S S, Zong W W, et al. Use of Tencent Street View Imagery for Visual Perception of Streets. *International Journal of Geo-Information*, 2017, 6(9): 265.
- [5] Carrasco-Hernandez R, Smedley A R D, Webb A R. Using urban canyon geometries obtained from Google Street View for atmospheric studies: Potential applications in the calculation of street level total shortwave irradiances. *Energy and Buildings*, 2015, 86: 340-348.
- [6] Dong R C, Zhang Y L, Zhao J Z. How Green Are the Streets Within the Sixth Ring Road of Beijing? An Analysis Based on Tencent Street View Pictures and the Green View Index. *International Journal of Environmental Research And Public Health*, 2018, 15(7): 1367.
- [7] Fry D, Mooney S J, Rodriguez D A, et al. Assessing Google Street View Image Availability in Latin American Cities. *Journal of Urban Health-Bulletin of the New York Academy of Medicine*, 2020, 97(4): 552-560.
- [8] Aikoh T, Homma R, Abe Y. Comparing Conventional Manual Measurement of the Green View Index with Modern Automatic Methods Using Google Street View and Semantic Segmentation. *Urban Forestry & Urban Greening*, 2023. DOI: 10.1016/j.ufug.2023.127845.
- [9] Zhao H B, Yu D F, Miao C H, et al. The Location Distribution Characteristics and Influencing Factors of Cultural Facilities in Zhengzhou Based on POI Data. *Scientia Geographica Sinica*, 2018, 38(9): 1525-1534.
- [10] Ta N, Zeng Y T, Zhu Q Y, et al. Relationship between built environment and urban vitality in Shanghai downtown area based on big data. *Scientia Geographica Sinica*, 2020, 40(1): 60-68.

- [11] Mu H K, Gao Y, Wang Z Y, et al. Equity Evaluation of Park Green Space Service Level from the Perspective of Supply and Demand Balance: An Empirical Analysis based on big Data. *Urban Development Studies*, 2019, 26 (11): 10-15.
- [12] Qi R H, Yang H, Wang S L, et al. Study on Evaluation and Planning of Urban Parks Based on Baidu POI Data. *Chinese Landscape Architecture*, 2018, 34(3): 32-37.
- [13] Carrasco-Hernandez, R, Smedley A R D, Webb A R. Using urban canyon geometries obtained from Google Street View for atmospheric studies: Potential applications in the calculation of street level total shortwave irradiances. *Energy And Buildings*, 2015, 86: 340-348. DOI: 10.1016/j.enbuild.2014.10.001.
- [14] Li X J, Ratti C. Mapping the spatial distribution of shade provision of street trees in Boston using Google Street View panoramas. *Urban Forestry & Urban Greening*, 2018, 31: 109-119. DOI: 10.1016/j.ufug.2018.02.013.
- [15] Cinnamon J, Gaffney A. Do-It-Yourself Street Views and the Urban Imaginary of Google Street View. *Journal of Urban Technology*, 2022, 29(3): 95-116. DOI: 10.1080/10630732.2021.1910467.
- [16] Wang X, Li Q, Guo Q, Wu H, et al. The generalization and construction of Voronoi diagram and its application on delimitating city' s affected coverage. *Journal Of Central China Normal University(Nat .Sci .)*, 2002, 36(01): 107-111. DOI: 10.19603/j.cnki.1000-1190.2002.01.027.

PERFORMANCE COMPARISON OF FREE-PISTON STIRLING CRYOCOOLERS WITH METALLIC AND NON-METALLIC PACKING IN WOUND REGENERATORS

ShuLing Guo, AnKuo Zhang*

Department of Refrigeration and Cryogenic Engineering, Shanghai Ocean University, Shanghai 201306, China.

Corresponding Author: AnKuo Zhang, Email: zhangkankuo@126.com

Abstract: In recent years, cryogenic technology has undergone continuous advancement. Compared with traditional multi-stage cascade vapor-compression cryocoolers, Stirling cryocoolers exhibit advantages such as compact structure, high reliability, and low environmental pollution, thus garnering significant attention from researchers. The regenerator, being the most costly component within Stirling cryocoolers, can see substantial production cost reductions through optimization of regenerator packing materials. Wound regenerators, owing to their compatibility with mechanical winding processes, eliminate material waste during fabrication. This approach significantly reduces both labor and material costs compared to stacked wire mesh regenerators, positioning wound configurations as a promising solution for cost-effective regenerator design. This study systematically investigates the internal losses, working characteristics, and refrigeration performance of a free-piston Stirling cryocooler prototype featuring wound regenerators. Experimental results demonstrate that at an operating temperature of 187 K, the Stirling cryocooler with non-metallic regenerator packing achieved a cooling capacity of 280 W, outperforming its metallic counterpart 180 W. Furthermore, the non-metallic variant exhibited a COP 0.2 higher than the metallic regenerator system, conclusively establishing the superior thermodynamic performance of non-metallic packing in wound regenerators.

Keywords: Heat transfer; Cryogenic; Stirling cryocooler; Cooling efficiency

1 INTRODUCTION

In recent years, cryogenic technology has continuously developed. Small Stirling cryocoolers have been widely applied in fields such as low-temperature storage. Compared with traditional multi-stage cascade cryocoolers, Stirling cryocoolers have more environmentally friendly working fluids and simpler, more efficient systems, thus being highly favored[1]. Among them, free-piston Stirling cryocoolers use linear motors to replace the traditional crankshaft-connecting rod drive structures and employ flexure springs and gas bearing technology for support, possessing advantages of low noise, small vibration, and higher compactness. Since their inception, they have received extensive attention from researchers[2-3].

Due to limitations in production costs and technical costs, Stirling cryocoolers were initially mainly used in aerospace and infrared detection fields. With the continuous maturation of Stirling refrigeration technology, the applications of Stirling cryocoolers have begun to expand. In the past decade, Stirling cryocoolers oriented toward general consumers have become a new direction of research. As a regenerative cryocooler, the regenerator is a critical component affecting the refrigeration performance of Stirling cryocoolers. The packing methods and parameters of regenerator packings have a direct impact on the regenerator's performance. Selecting appropriate packing materials is crucial for the overall system performance of the cryocooler[4]. Meanwhile, the regenerator is also one of the most expensive components in Stirling cryocoolers. Therefore, to reduce the manufacturing costs of Stirling cryocoolers, optimization of regenerator packing can be implemented[5]. Currently, stacked wire-mesh is the most commonly used packing method for Stirling cryocooler regenerators. This type of regenerator possesses high specific surface area and volumetric heat capacity, along with low axial thermal conductivity. However, it suffers from disadvantages such as metal material waste and complex manufacturing processes. The production process of stacked wire-mesh regenerators requires cutting square wire-mesh into annular components needed for packing, which wastes significant material. Additionally, manually packing the wire-mesh into the regenerator substantially increases labor costs. In contrast, wound wire-mesh structures can utilize automated mechanical winding processes, enabling drastic reductions in both labor costs and material costs, thereby serving as an effective solution for lowering regenerator costs.

Currently, research targeting wound wire mesh regenerators remains relatively limited. Wound wire mesh can be categorized into metallic and non-metallic types based on material differences. Metallic wire-mesh is typically fabricated from stainless steel. The Technical Institute of Physics and Chemistry, Chinese Academy of Sciences experimentally compared spiral-wound metallic wire-mesh with traditional stacked metallic wire-mesh. Results demonstrated that compared to stacked wire-mesh, spiral-wound configurations achieved over 80% reduction in both manufacturing costs and processing time. While stacked wire-mesh regenerator cryocoolers exhibit superior refrigeration efficiency, spiral-wound variants demonstrate lower flow resistance with efficiency comparable to vapor-compression refrigeration systems, showing promising application potential in deep-cryogenic cooling performance. Non-metallic wire-mesh is generally fabricated by winding polyester films[6]. Zhu Haifeng et al. investigated the use of PET fibers as a metallic substitute for regenerator packing, with the input PV power increasing

by only 5W at 1W@80K operating conditions, demonstrating the feasibility of replacing metallic materials with non-metallic alternatives in regenerator packing[7]. Cui Yunhao et al. conducted comparative performance tests on regenerators employing random stainless steel wire-mesh packing versus polyimide film spiral-wound packing. Results demonstrated that the spiral-wound non-metallic wire-mesh regenerator exhibited a 16% higher relative Carnot efficiency compared with the randomly packed stainless steel wire-mesh regenerator[8].

Wound wire mesh regenerators exhibit distinct performance variations in cryocoolers depending on their material composition. This study investigates the operational differences in cryocoolers caused by metallic versus non-metallic materials under identical packing configurations through numerical simulations and experimental validation. Based on a prototype free-piston Stirling cryocooler, a comprehensive Sage model of the entire refrigeration system was established. Comparative analyses were conducted on temperature distributions, internal loss characteristics, and their spatial profiles across regenerators with different materials. The research systematically examines the impacts of packing parameters and operational variables on refrigeration efficiency, while determining optimal packing specifications, operating frequencies, and charge pressures. Predictive simulations and performance comparisons were executed for both metallic and non-metallic regenerator configurations.

2 NUMERICAL CALCULATION

To investigate the performance differences between free-piston Stirling cryocoolers with metallic versus non-metallic spiral-wound regenerators, this study employs the one-dimensional cryocooler modeling software Sage for computational analysis. The software numerically solves the governing equations of mass, energy, and momentum conservation through finite difference method to obtain simulation results. Sage utilizes graphical interface-based modeling, constructing system-level simulations by defining dimensional parameters, geometric configurations, and material properties of individual cryocooler components. These components are interconnected through mass flow dynamics, pressure wave propagation, mechanical forces, and energy transfer pathways, thus enabling comprehensive system-level simulation and performance optimization.

2.1 Thermodynamic Analysis of Free-Piston Stirling Cryocooler

The free piston Stirling cooler is a closed system. Under the drive of the motor, the input power obtained by the compression piston is \dot{W} ; gas working fluid releases heat to the external environment, with a heat release of \dot{Q}_1 ; the gas working fluid drives the movement of the exhaust device, and the amount of work done to the exhaust device is \dot{W}_0 ; at the expansion chamber, the gas working fluid absorbs heat from the cold source, producing a cooling capacity of \dot{Q}_0 . The external ambient temperature is T_1 . The cold end temperature is T_0 .

For a free piston Stirling cooler, the first and second laws of thermodynamics can be expressed as:

$$\dot{W} - \dot{W}_0 = \dot{Q}_1 - \dot{Q}_0 + \frac{d(mu)}{dt} \quad (1)$$

$$\frac{\dot{Q}_1}{T_1} - \frac{\dot{Q}_0}{T_0} + \dot{S}_{ex} = \frac{d(mt)}{dt} \quad (2)$$

where, $\frac{d(mu)}{dt}$ is the rate of change of internal energy in the refrigeration system, $\frac{d(mt)}{dt}$ is The rate of change of entropy in the refrigeration system is 0 in steady state.

By combining the above equations, the COP of the refrigeration unit can be obtained as[9]:

$$\text{COP} = \frac{\dot{Q}_0}{\dot{W} - \dot{W}_0} = \frac{T_0}{T_1 - T_0} \left[1 - \frac{T_0 \dot{S}_{ex}}{\dot{W} - \dot{W}_0} \right] \quad (3)$$

2.2 Numerical Calculation

Within the numerical model, the one-dimensional governing equations of momentum, continuity, and energy within the gas domain are as follows:

$$\frac{\partial \rho A}{\partial t} + \frac{\partial \rho u A}{\partial x} = 0 \quad (4)$$

$$\frac{\partial \rho u A}{\partial t} + \frac{\partial \rho u^2 A}{\partial x} + \frac{\partial P}{\partial x} A - F_A = 0 \quad (5)$$

$$\frac{\partial \rho e A}{\partial t} + P \frac{\partial A}{\partial t} + \frac{\partial}{\partial x} (u \rho e A + u P A + q) - \dot{Q}_w = 0 \quad (6)$$

where P , u , A and ρ denote the working gas pressure, mean-flow velocity in the x direction (longitudinal), cross sectional area and working gas density, respectively[10].

Due to non-ideal factors, there are a series of losses within the FPSC that affect its refrigeration efficiency. The losses of the cryocooler are divided into static losses and dynamic losses. Static loss is mainly heat conduction loss.

The axial heat conduction loss caused by heat conduction can be solved based on Fourier's law:

$$Q_{\text{con}} = -\lambda A \frac{dT}{dx} \quad (7)$$

Dynamic losses are typically the most significant losses during the operation of a cryocooler. Dynamic losses of a cryocooler include incomplete heat exchange losses, pressure drop losses, and leakage losses.

In an ideal regenerator, heat exchange between the packings and the working fluid is highly efficient. However, in actual operation, due to the temperature difference between the gas flow and the regenerator, there is an insufficient heat exchange phenomenon, resulting in non-ideal heat transfer losses. The non-ideal heat transfer losses Q_r can be calculated as follows:

$$Q_r = \dot{m}_r c_p (1 - \eta)(T_{cr} - T_{er}) \quad (8)$$

where \dot{m}_r is the average mass flow rate of gas through the regenerator during the cold and hot blow periods, c_p is the average specific heat capacity at constant pressure of the working fluid, and T_{cr} , T_{er} are the temperature at the cold and hot ends of the regenerator.

The flow resistance inherent in cryocooler operation induces amplitude attenuation of the working fluid's oscillatory flow, consequently generating flow resistance losses. The definition of pressure drop loss Q_f is as follows:

$$Q_f = \oint \Delta P_r dV_e \quad (9)$$

where ΔP_r is the pressure drop across the regenerator, and V_e is the volume of the expansion volume.

Flow resistance losses mainly occurs in the regenerator part of the cryocooler. The pressure drop ΔP_r in the regenerator is as follows:

$$\Delta P_r = \frac{f_r G_r^2 L_r}{2K_{ci} R_{Mr} \rho_{Mr}} \quad (10)$$

where C_f is the friction factor, G_r is the mass flow rate per unit flow area, L_r is the length of the flow path, K_{ci} is the unit conversion coefficient, and R_{Mr} is the hydraulic radius of the flow path.

2.2 Regenerator Material Property Comparison

The regenerator is the core component enabling work-heat conversion. Alternating fluid and solid packing continuously undergo heat exchange within the regenerator, thus the selection of regenerator packing must satisfy thermophysical property requirements. The volumetric heat capacity of regenerator packing should be significantly greater than that of the working gas[2]; large heat transfer area ensures sufficient heat exchange between gas and packing; low flow resistance reduces working fluid flow losses; small axial thermal conductivity decreases cold-end heat losses. While packing structures simultaneously satisfying all these characteristics are difficult to achieve, appropriate regenerator packing can be selected based on specific application requirements. Among them, spiral-wound regenerators feature simple structures and lower manufacturing costs. The structural diagram and physical diagram of the spiral-wound regenerator are shown in Figure 1.

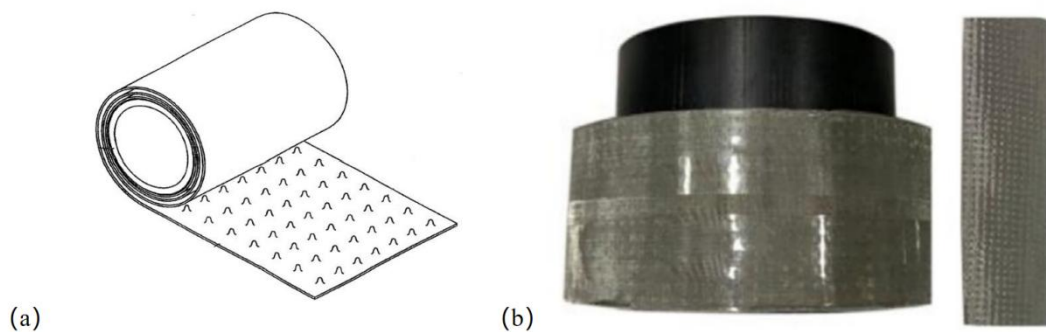


Figure 1 Non-Metallic wound Regenerator Structural Diagram (a) and Physical Diagram of Metallic wound Regenerator Packing (b) [6]

Wound regenerator packing is categorized into metallic and non-metallic types. Metallic packing typically employs 304 stainless steel. Non-metallic packing commonly utilizes polyester materials such as polyester (PET), PEN, Teflon, and polyimide. Due to material properties, stainless steel regenerators exhibit significantly higher axial thermal conductivity than non-metallic materials, resulting in greater axial heat losses within the regenerator. However, stainless steel demonstrates superior volumetric specific heat capacity compared to non-metallic materials, endowing it with enhanced thermal energy storage capabilities. At 100K, the volumetric specific heat of helium is approximately 0.86 J/(m³·K), significantly lower than that of both regenerator packing materials. Therefore, both material types can be effectively utilized for heat exchange in regenerators (Figure 2).

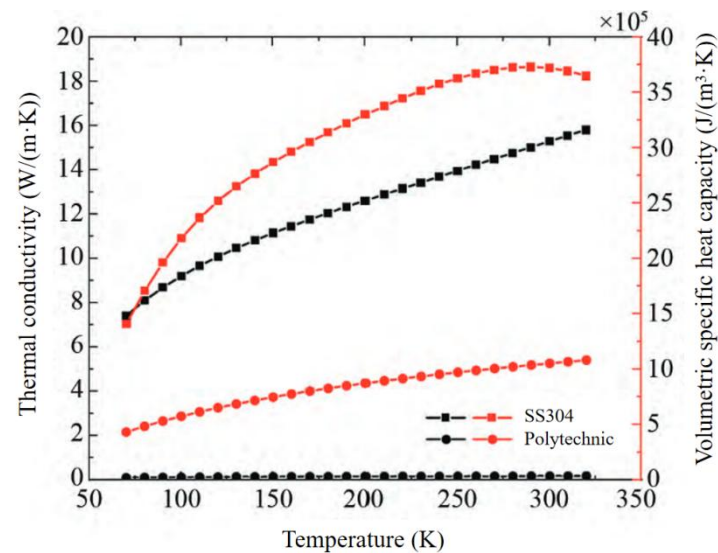


Figure 2 Comparison of Thermal Conductivity and Volumetric Heat Capacity between Polyimide and 304 Stainless Steel as Functions of Temperature[8]

The spiral-wound metallic packing is typically fabricated by winding metallic wire mesh, featuring a more porous and loose structure that provides a larger specific surface area. In contrast, non-metallic polyester films usually employ perforated surfaces, resulting in relatively lower porosity and smaller specific surface area. In Stirling cryocoolers, the regenerator generally reciprocates with the displacer within the expansion cylinder. A greater mass leads to increased mechanical vibration during operation, necessitating higher radial stiffness in the flexure springs[8]. Excessive vibration and friction can also reduce the operational lifespan of the cryocooler. Non-metallic materials, being lighter than their metallic counterparts, can decrease the regenerator mass and mitigate vibration-related issues. The flow channels in spiral-wound regenerators are parallel and well-ordered, ensuring relatively low flow resistance for both metallic and non-metallic materials.

Compared with metallic packing, non-metallic packing has significantly lower costs. Taking PEN material as an example, the material cost for spiral-wound PEN is only 5% of that for 200-mesh wire mesh, with processing and filling costs also being lower than those for metallic wire mesh. The total production cost of PEN material regenerators is approximately 16% of that for metallic materials (Table 1).

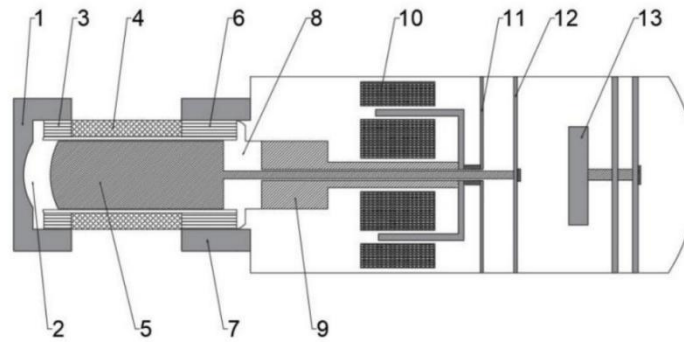
Table 1 Total Cost Estimation of Metallic and Non-Metallic Packing for Wound Regenerators

Material	Wound PEN material	Wound 200 mesh metal wire mesh
Raw material cost/USD	0.7	13.8
Processing and filling costs/USD	2.1	2.8
Total cost/USD	2.8	16.6

3 SIMULATION SETUP

To further analyze the differences between metallic and non-metallic packing regenerators, this study employs one-dimensional simulation software to model two distinct cryocooler configurations. Concurrently, an experimental platform was established to validate both the simulation results and the actual performance of the two cryocooler types. This section details the model configuration and experimental apparatus specifications.

The schematic diagram of the prototype used for experimental research in this study is shown in the figure below. This free-piston Stirling cryocooler primarily consists of the following components: cold head, cold-end heat exchanger, regenerator, hot-end heat exchanger, displacer piston, compression piston, linear motor, flexure springs, and vibration damping device. The low-temperature cooler generates an alternating magnetic field by using a high-frequency AC power source to drive the reciprocating motion of the compression piston. This movement forces the helium working fluid to produce a cooling effect through the cold fingers. In addition, a vibration reduction device is installed at one end of the system to effectively reduce vibration and noise. For the metallic packing regenerator, it is fabricated by spirally winding strip-shaped SS304 stainless steel wire mesh, with porosity adjusted through mesh protrusions. The non-metallic packing utilizes PEN film, which is first perforated on its surface and then wound into shape (Figure 3).



1.cold head; 2.expansion volume; 3.cold heat exchanger; 4.regenerator; 5.displacer; 6.hot heat exchanger; 7.hot heat radiator; 8.compression volume; 9.compression piston; 10.linear motor; 11.compression flexure bearing; 12.displacer flexure bearing; 13.vibration damping device

Figure 3 Structural Diagram of Free-Piston Stirling Cryocooler

In the model, the cryocooler operates at a working frequency of 55 Hz with a charge pressure of 3.0 MPa, using helium as the refrigerant. The external environment of the refrigerant is set to 310 K. Since the cryocooler primarily serves as the cold source for a low-temperature refrigerator, the cold-end temperature is set to 187 K.

4 RESULTS AND DISCUSSION

4.1 Axial Temperature Distribution along the Entire Cryocooler with Metallic Versus Non-Metallic Packing Regenerators

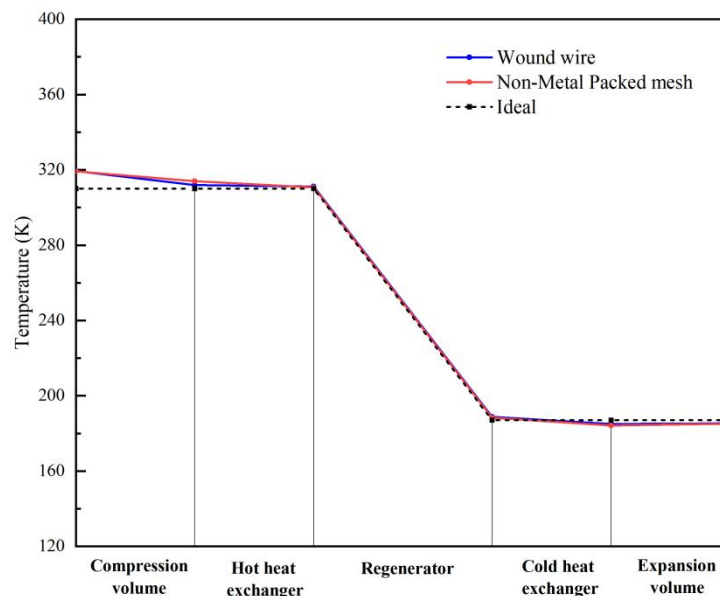


Figure 4 Axial Temperature Distribution Diagram of Complete Cryocooler with Metallic vs. Non-Metallic Packing

Figure 4 shows the axial temperature distribution from the compression chamber inlet to the expansion chamber outlet comparing the actual internal temperatures of Stirling cryocoolers with metallic and non-metallic regenerators against the theoretical temperature distribution. In the compression chamber, the actual internal temperature of the cryocooler is higher than the ideal temperature. This occurs because the ideal cryocooler process assumes isothermal compression, whereas the actual cryocooler undergoes near-adiabatic compression, resulting in significantly higher gas temperatures at the compression chamber end compared to theoretical values. From the compressor to the hot-end heat exchanger outlet, the cryocooler temperature continuously decreases. Before the gas reaches the hot end of the regenerator, its temperature remains higher than the ideal temperature, with the non-metallic regenerator cryocooler showing even higher temperatures in the compression chamber. In the hot-end heat exchanger, the temperature decrease rate slows down. Within the regenerator, the temperature drops rapidly, with both cryocooler types showing good agreement with the ideal temperature drop rate. At the cold-end heat exchanger, the actual gas temperature decreases slightly, with both cryocooler types exhibiting similar temperatures. From the expansion chamber inlet to outlet, the actual cryocooler

temperature rises slightly because the working gas absorbs heat during expansion, causing a minor temperature drop at the expansion chamber outlet. Both cryocooler configurations show essentially identical temperature distributions from the cold-end heat exchanger inlet to the expansion chamber outlet.

4.2 Impact of Regenerator Parameters on Cooling Efficiency in Cryocoolers with Metallic Versus Non-Metallic Packing

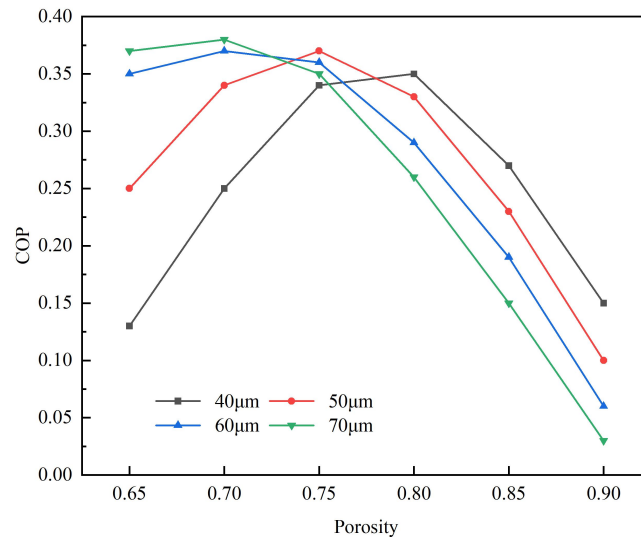


Figure 5 The Relationship between Porosity and COP of Wound Wire Mesh Stirling Cryocooler Regenerators at Different Wire Diameters

For metallic regenerators, the wire diameter of the metal mesh and the porosity of the packing are the main regenerator parameters. Figure 5 shows the relationship between porosity and COP for metallic regenerators with different wire diameters. In the 40-70 μm wire diameter range, COP first increases and then decreases as porosity increases. This is because when the wire diameter is fixed, at lower porosity levels, increasing porosity creates more voids between the metal materials, allowing smoother gas flow and reduced flow resistance while maintaining sufficient effective heat transfer area. This enables more efficient heat exchange between the gas and metal materials, improving the regenerator's heat transfer performance and thereby increasing the cryocooler's COP. However, when porosity exceeds a certain value, the proportion of metal material decreases, meaning the effective surface area available for heat exchange is reduced. With fewer contact opportunities between the gas and metal materials, heat exchange becomes insufficient, incomplete heat transfer losses increase, leading to degraded cryocooler performance and lower COP.

The porosity corresponding to maximum COP shifts toward lower values as wire diameter increases. This occurs because larger wire diameters result in greater actual metal volume fraction within the same space. Consequently, sufficient metal surface area for heat exchange can be maintained even at lower porosity levels. Compared with fine wires, excessively high porosity becomes unnecessary for preserving effective heat transfer area; simultaneously, thicker wires make gas flow channels relatively narrower and more complex within the regenerator. At higher porosity, although gas flow space increases, the coarser wires cause more significant flow resistance due to enhanced flow disturbances and friction in the channels. Therefore, for thicker wires, relatively lower porosity optimizes gas flow conditions, achieving better balance between flow resistance and heat transfer effectiveness, thus maximizing COP at reduced porosity levels. The maximum COP occurs at 70 μm wire diameter. Peak COP values appear at 70 μm /0.7 porosity, followed by 60 μm /0.7 and 50 μm /0.75 combinations.

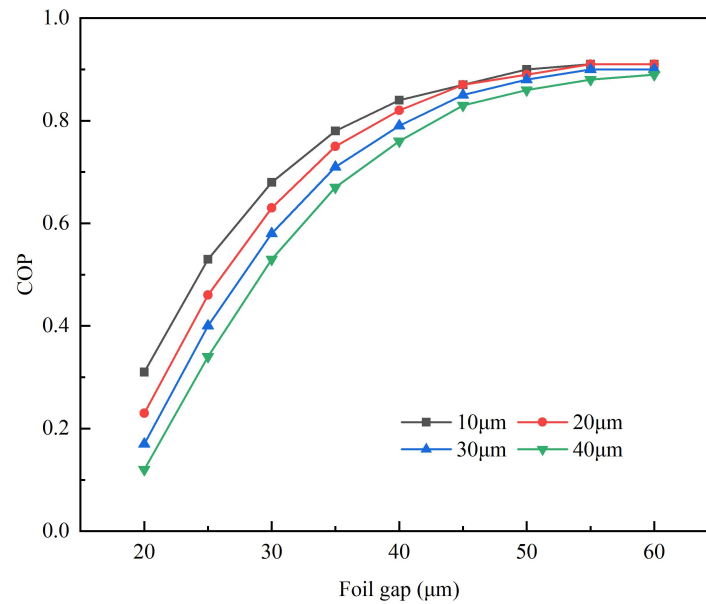


Figure 6 The Relationship between the Foil Gap and the COP of the Regenerator in a Wound Non-Metallic Stirling Cryocooler under Different Foil Thicknesses

The primary regenerator parameters for non-metallic regenerators are foil thickness and foil gap. Figure 6 presents the relationship between foil gap and COP for non-metallic regenerators under varying foil thicknesses. Unlike metallic regenerators, within the 10-40 μm foil thickness range, COP increases with growing foil gap until stabilizing after reaching 45 μm gap. At smaller gaps, confined gas flow channels result in significant flow resistance losses that impair cooling efficiency. Consequently, COP shows marked improvement with increased gap. However, excessive gap reduces gas-packing heat exchange effectiveness, causing substantial incomplete heat transfer losses. When gap exceeds 45 μm , these losses outweigh flow resistance effects on cooling efficiency, leading to stabilized COP growth. Maximum COP values consistently occur at 55 μm gap across all thicknesses.

Within the 10 μm -40 μm foil gap range, the influence of increasing foil thickness on COP remains relatively minor. As foil thickness grows, the maximum COP shows slight decrease, though all configurations achieve similar peak COP values. The absolute maximum COP occurs at 20 μm foil thickness with 55 μm gap.

4.3 Impact of Working Parameters on Cooling Efficiency in Cryocoolers with Metallic Versus Non-Metallic Packing Regenerators

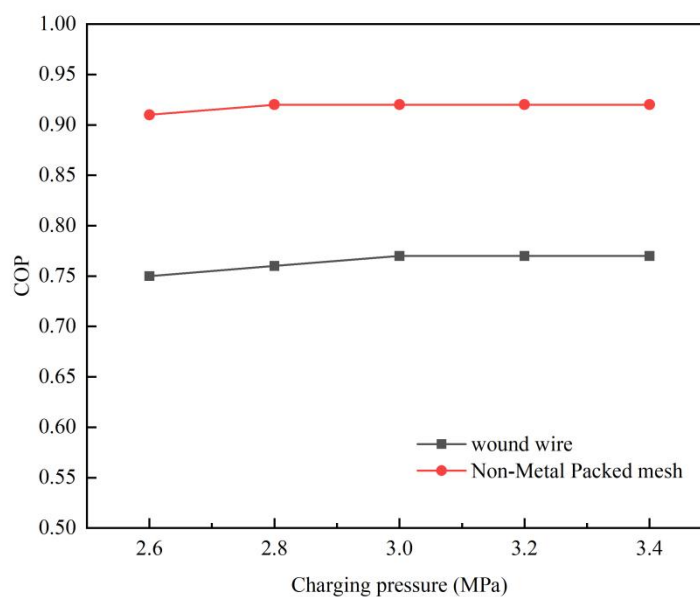


Figure 7 Comparison of Charging Pressure and COP in Metallic and Non-Metallic Cryocoolers

Figure 7 shows the variation of Stirling refrigerators with different material recuperators under different inflation

pressures. COP increases with rising charge pressure, exhibiting a higher growth rate below 3.0 MPa before stabilizing beyond this point. The maximum COP of 0.77 is attained within the 3.0-3.4 MPa range, with no significant further increase thereafter. This occurs because increased charge pressure enables higher compressor input power at identical piston stroke while simultaneously increasing working gas density, which enhances volumetric heat capacity and refrigeration capacity per cycle during expansion, thereby boosting PV work, theoretical refrigeration capacity and COP. However, as charge pressure continues to rise, refrigeration losses progressively increase. When the incremental gain in theoretical refrigeration capacity approaches the total loss increase, COP stabilizes at its maximum value. At charge pressures below 2.6 MPa, the lower COP indicates severely constrained cryocooler performance due to insufficient pressure.

The observed COP variation pattern primarily stems from elevated charge pressure increasing gas density, which alters heat transfer characteristics between the gas and regenerator packing under isothermal conditions. This exacerbates insufficient heat transfer, resulting in inadequate thermal exchange that amplifies incomplete heat transfer losses. Concurrently, the increased gas density and viscosity under higher charge pressure amplify flow resistance within the regenerator channels, intensifying flow resistance losses. Thermal conduction losses show minimal variation compared to flow resistance and incomplete heat transfer losses.

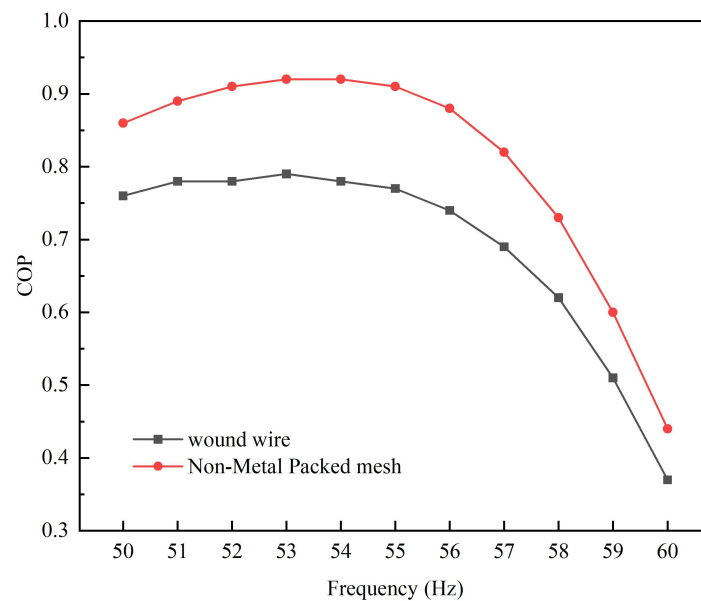


Figure 8 Comparison of the Relationship between the Working Frequency and COP of the Metallic and Non-Metallic Packing Regenerator Cryocoolers

Figure 8 presents a comparison of the relationship between operating frequency and COP for metallic and non-metallic regenerator cryocoolers at a charge pressure of 3.0 MPa. Simulations reveal that both metallic and non-metallic cryocoolers exhibit similar variation patterns in cooling efficiency with frequency, showing an initial increase followed by a decrease, with peak COP occurring at 55 Hz. The non-metallic regenerator consistently demonstrates higher COP values than its metallic counterpart, indicating superior refrigeration performance.

The variation in COP originates from disparities in thermal penetration efficiency. At excessively low frequencies, insufficient thermal penetration reduces heat exchange efficiency. Higher frequencies shorten the refrigeration cycle duration, accelerating gas flow velocity through the regenerator. This increases gas-packing contact frequency per unit time; although individual contact duration may decrease, the cumulative heat transfer opportunity rises, improving heat exchange completeness and reducing corresponding losses. Concurrently, elevated frequency transitions flow regimes from laminar toward turbulent conditions, where enhanced fluid mixing reduces boundary layer thickness and dominates over increased dynamic pressure effects, ultimately decreasing flow resistance losses. Axial conduction loss remains stable as it primarily depends on axial temperature gradients, which show minimal sensitivity to frequency variations.

4.4 Comparison of Refrigeration Losses between Metallic-Packing and Non-Metallic Packing Regenerator Cryocoolers

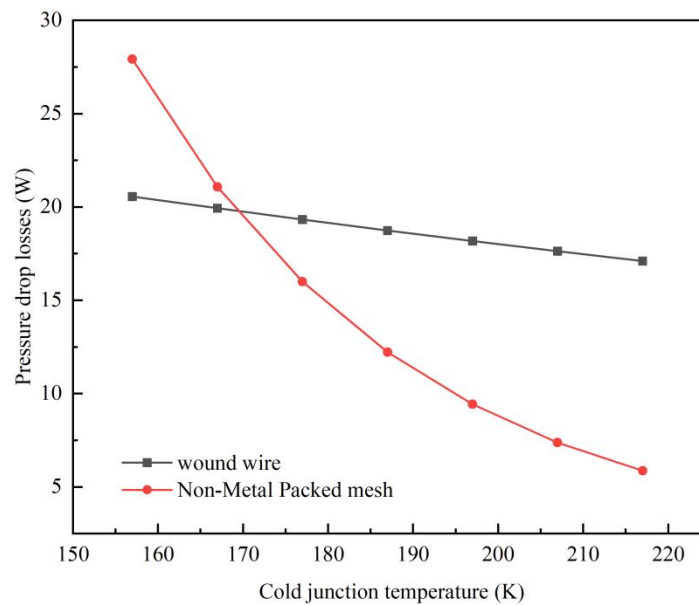


Figure 9 Comparison of Flow Resistance Losses in Stirling Cryocoolers with Metallic and Non - Metallic Regenerators at Different Temperatures

Figure 9 presents a comparison of flow resistance losses between metallic and non-metallic regenerator cryocoolers at different temperatures. The results show that as the cold-end temperature increases from 150K to 220K, the flow resistance losses of both metallic and non-metallic regenerator cryocoolers exhibit a decreasing trend. This is because at higher gas temperatures, the viscosity and density of the gas are relatively smaller, thereby reducing flow resistance. At lower cold-end temperatures, due to higher gas viscosity, both metallic and non-metallic regenerators show larger flow resistance losses. When the cold-end temperature is below 173K, the flow resistance of non-metallic regenerators exceeds that of metallic regenerators. At 160K, the flow resistance loss of the non-metallic material is 27W, while that of the metallic material is 20W. However, as the cold-end temperature rises, the flow resistance loss of the non-metallic material decreases more rapidly. When the cold-end temperature exceeds approximately 170K, its flow resistance loss becomes lower than that of the metallic material. At 220K, the flow resistance loss of the non-metallic material drops to about 6W, whereas the metallic material remains at approximately 18W. When the temperature exceeds 173K, the metallic surfaces exhibit higher roughness, while non-metallic materials typically have lighter weight and smoother surfaces, resulting in relatively smaller flow resistance and lower flow resistance losses. In conclusion, non-metallic regenerators demonstrate smaller internal losses and superior regenerative performance compared to metallic regenerators.

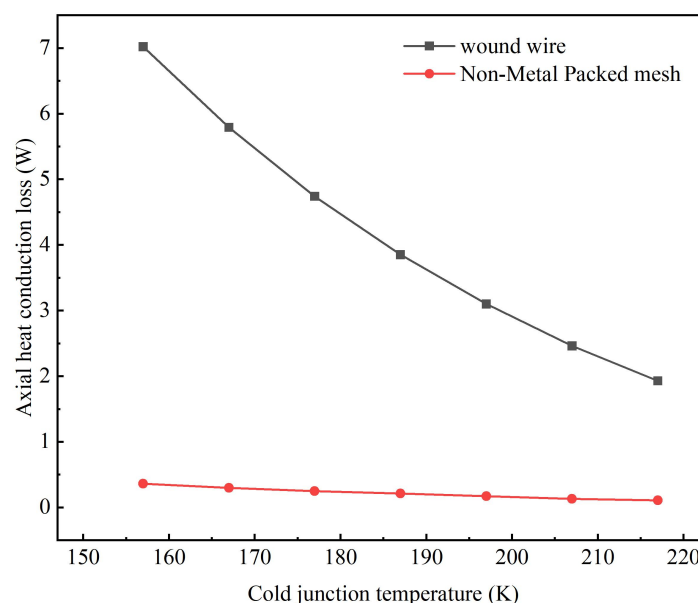


Figure 10 Comparison of Axial Heat Conduction Losses in Stirling Cryocoolers with Metallic And Non - Metallic

Regenerators at Different Temperatures

Figure 10 presents a comparison of axial conduction losses between metallic and non-metallic regenerator cryocoolers at different temperatures. The results show that as the cold-end temperature increases from 150K to 220K, the axial conduction loss of metallic regenerators decreases significantly, while that of non-metallic regenerators shows negligible variation. The reduction in metallic regenerator conduction loss is related to the decreased axial temperature gradient caused by the rising cold-end temperature. From the figure, it is evident that the axial conduction loss of metallic regenerators is significantly greater than that of non-metallic regenerators. During the operation of Stirling cryocoolers, distinct temperature gradients exist within the regenerators. Metallic materials inherently possess higher thermal conductivity compared to many non-metallic materials, enabling more efficient heat conduction. Additionally, since the spiral-wound wire mesh regenerator packing is integrally formed, there are no air gaps between metal wires as in stacked metal mesh configurations, which further exacerbates axial conduction. In the regenerator, when temperature gradients exist, the metal wire mesh rapidly transfers heat from high-temperature regions to low-temperature regions. Excessive axial conduction in regenerators adversely affects the performance of free-piston Stirling cryocoolers in multiple ways: Excessive axial conduction leads to excessive heat transfer along the axial direction, disrupting the temperature distribution between the hot and cold ends and reducing the regenerator's heat recovery efficiency. Furthermore, due to high axial conduction, the cryocooler requires additional energy consumption to maintain the low temperature at the cold end. This extra energy is not utilized for effective refrigeration but is wasted in compensating for the cold-end temperature rise caused by axial conduction, thereby negatively impacting refrigeration efficiency. Therefore, compared to non-metallic materials, the high axial conduction loss of metallic materials constitutes one of the primary defects of spiral-wound metallic regenerators.

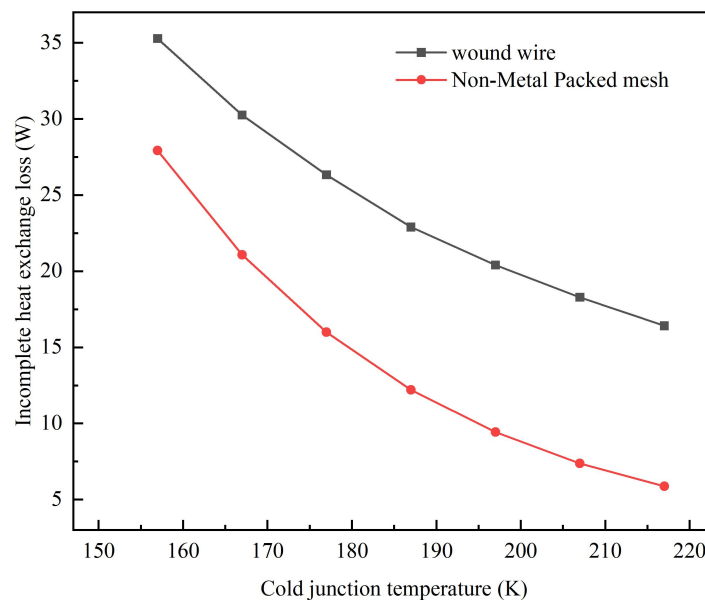


Figure 11 Comparison of Incomplete Heat Transfer Losses in Stirling Cryocoolers with Metallic and non - Metallic Regenerators at Different Temperatures

Figure 11 presents a comparison of incomplete heat transfer losses between metallic and non-metallic regenerator cryocoolers at different temperatures. The results show that as the cold-end temperature rises from 150K to 220K, both metallic and non-metallic regenerator cryocoolers exhibit decreasing trends in incomplete heat transfer losses. This reduction correlates with the diminished temperature difference between the working gas and regenerator materials under elevated temperatures, which enhances heat exchange efficiency. At 150K, the incomplete heat transfer loss of the metallic regenerator measures 35W, compared to 28W for the non-metallic regenerator. When the cold-end temperature reaches 220K, these losses decrease to 17W and 6W for metallic and non-metallic regenerators respectively. Under identical cold-end temperatures, metallic regenerators consistently demonstrate higher incomplete heat transfer losses than non-metallic counterparts. The reduction magnitude of incomplete heat transfer losses in metallic regenerators is relatively smaller compared to non-metallic regenerators. The elevated incomplete heat transfer loss in metallic regenerators is directly related to their high axial conduction loss. The axial heat transfer disrupts the normal temperature distribution within the regenerator, leading to reduced temperature gradients during subsequent heat exchange processes. This diminished thermal gradient lowers heat transfer efficiency and thereby increases incomplete heat transfer losses.

4.4 Comparison of Refrigeration Performance between Metallic Packing and Non-Metallic Packing Regenerator Cryocoolers

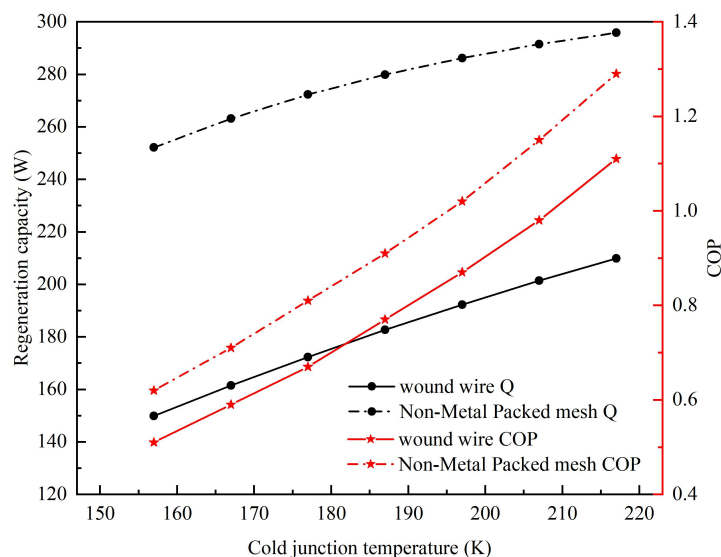


Figure 12 The Cooling Capacity and COP of the Cryocoolers with Metallic and Non-Metallic Regenerator under Different Cold End Temperatures

Figure 12 illustrates the refrigeration performance differences between metallic and non-metallic regenerator cryocoolers at various cold-end temperatures. At 187K operating temperature, the non-metallic regenerator cryocooler achieves 280W refrigeration capacity, whereas its metallic counterpart delivers only 180W under identical conditions. Concurrently, the non-metallic regenerator exhibits 0.2 higher COP and superior cooling efficiency, demonstrating better overall refrigeration performance. Comprehensive analysis confirms the non-metallic regenerator cryocooler outperforms the metallic version. The spiral-wound regenerator Stirling cryocooler maintains refrigeration capacity exceeding 150W with COP above 0.7 across the 180K temperature range, showing comparable efficiency to stacked metal mesh configurations at equivalent operating temperatures, making it an ideal cost-effective regenerator packing solution.

5 CONCLUSION

This chapter conducts numerical simulations based on a free-piston Stirling cryocooler prototype. A Sage model for the spiral-wound regenerator free-piston Stirling cryocooler was established with defined boundary conditions. Comparative simulations were performed on temperature field distribution, energy flow patterns, loss mechanisms, operational characteristics, and cooling efficiency between two regenerator material configurations. The full-scale Sage numerical model was developed based on the prototype, with parameter adjustments for metallic and non-metallic regenerators. Optimal packing parameters were determined through simulations: metallic regenerators achieved peak performance at 0.70 porosity with 70 μ m wire diameter, while non-metallic regenerators optimized at 55 μ m flow channel spacing and 20 μ m film thickness. Both configurations exhibited similar parametric trends, sharing optimal operating frequency (55Hz) and charge pressure (3.0MPa). Internal component losses increased with elevated charge pressure, though insufficient pressure (<2.6MPa) severely constrained cooling efficiency. While higher operating frequencies reduced internal losses, decreased displacer amplitude degraded refrigeration performance. At 187K operating temperature, the non-metallic regenerator delivered 280W refrigeration capacity versus 180W for the metallic counterpart, with a 0.2 higher COP and superior efficiency. Both configurations maintained COP values exceeding 0.7 across the operational temperature range. The spiral-wound regenerator design demonstrates promising refrigeration performance, providing new insights for developing cost-effective, high-efficiency Stirling cryocoolers with large cooling capacities.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCE

- [1] Getie M Z, Lanzetta F, Bégot S, et al. Reversed regenerative Stirling cycle machine for refrigeration application: A review. *International Journal of Refrigeration*, 2020, 118: 173–187. DOI: 10.1016/j.ijrefrig.2020.06.007.
- [2] Wang B, Xu H, Zhang W Q, et al. Low-Temperature cryocoolers Based on Gas-Bearing Stirling Cryocoolers. *Low Temperature and Superconductivity*, 2017, 45(8): 26-31. DOI: 10.16711/j.1001-7100.2017.08.005.
- [3] Liu S, Jiang Z, Ding L, et al, Impact of Operating Parameters on 80 K Pulse Tube Cryocoolers for Space Applications. *International journal of refrigeration*, 2019, 99: 226–233. DOI: 10.1016/j.ijrefrig.2018.12.026.
- [4] Costa S C, Barreno I, Tutar M, et al. The thermal non-equilibrium porous media modelling for CFD study of woven wire matrix of a Stirling regenerator. *Energy conversion and management*, 2015, 89: 473-483. DOI: 10.1016/j.enconman.2014.10.019.
- [5] Kim, Hong Seok, In Cheol Gwak, and Seong Hyuk Lee. Numerical Analysis of Heat Transfer Area Effect on Cooling Performance in Regenerator of Free-Piston Stirling Cooler. *Case studies in thermal engineering*, 2022, 32: 101875. DOI: 10.1016/j.csite.2022.101875.
- [6] Cui Y, Qiao J, Song B, et al. Experimental study of a free piston Stirling cooler with wound wire mesh regenerator. *Energy (Oxford)*, 2021, 23: 121287. DOI: 10.1016/j.energy.2021.121287.
- [7] Zhu Haifeng, Yinong Wu, Na Li, et al. Experimental Study on Novel Regenerator Packing for Stirling Refrigerators. *Proceedings of the 2015 Annual Academic Conference of the Shanghai Association of Refrigeration*, Shanghai Association of Refrigeration, 2015.
- [8] Cui Yunhao, Xiaotao Wang, Yanan Wang, et al. Simulation Study on the Application of Low-Cost Regenerator Packing in Liquid Nitrogen Temperature Range Stirling Refrigerators. *Vacuum and Cryogenics*, 2022, 28(3): 317–323. DOI: 10.3969/j.issn.1006-7086.2022.03.011.
- [9] Cai Yachao. Study on Operational Characteristics of High-Capacity Integral Stirling Refrigerators. PhD dissertation, Zhejiang University, 2015.
- [10] Gedeon D. Sage User's Guide: Stirling, Pulse-Tube and Low-T Cooler Model Classes. Gedeon Associates, 2014. [https://refhub.elsevier.com/S1359-4311\(19\)38873-8/h0230](https://refhub.elsevier.com/S1359-4311(19)38873-8/h0230).

