Analysis of the Current State of City Information Model Research and Construction Based on Data Mining

Zhiyuan Guo[1,3,a], Yue Kong[1,3,b], Xirui Cheng[4,c], Hanbin Luo[1,3,*], Zhenyuan Liu[2,d]

[1.]School of Civil and Hydraulic Engineering, Huazhong University of Science and Technology, Wuhan, 430000 China
[2]School of Artificial Intelligence and Automation Hust, Huazhong University of Science and Technology, Wuhan, 430000 China
[3]National Center for Technology Innovation for Digital Construction, Wuhan, 430000 China
[4]Wuhan Digital Construction Industry Technology Research Institute Co., Ltd., Wuhan, 430000 China

[a]guozhiyuan@hust.edu.cn; [b]kongyue1013@hust.edu.cn; [c]1037846298@qq.com; *luohbcem@hust.edu.cn; [d]zyliu@hust.eu.cn;

Abstract. Since the concept of City Information Modeling (CIM) was introduced, scholars in China have conducted extensive research and discussions on CIM, with the government also allocating significant resources to its development, achieving a breakthrough from non-existence to existence in CIM research and construction. To explore the current status and issues of CIM research and construction in China, this paper utilizes CiteSpace software and data mining methods, with data sourced from the CNKI database and public tender announcements on the internet. It systematically analyzes two aspects of CIM research and CIM construction status, and finds that CIM-related research and construction are in the beginning stage, and both show rapid development trend, there are problems such as focusing on the construction field, unbalanced construction in various regions, and large capital investment.

Keywords: City information model, Transformation of urbanization, Smart city construction, Urban construction.

1 Introduction

As the urbanization rate in China continues to climb, urban governance models are facing unprecedented challenges. By the end of 2023, the urbanization rate of the permanent population in China reached 66.16%. To address the difficulties in infrastructure planning and management, urban spatial governance, optimization of public services, and environmental protection and sustainable development brought about by increased urbanization, government departments need to consider how governance models can transition towards more refined, intelligent, and dynamic forms, achieving leapfrog development in the construction of smart cities[1]. In this context, the technology

F. Ding et al. (eds.), *Proceedings of the 2024 International Conference on Civil Engineering Structures and Concrete Materials (CESCM 2024)*, Advances in Engineering Research 247,
https://doi.org/10.2991/978-94-6463-564-5_6

of City Information Modeling (CIM) has emerged. Based on technologies such as Building Information Modeling (BIM), Geographic Information Systems (GIS), and the Internet of Things (IoT), CIM integrates multidimensional and multiscale spatial data of cities' above and below ground, indoors and outdoors, and historical to future scenarios, along with IoT perception data, to construct a three-dimensional digital spatial city information synthesis.

In November 2018, the Ministry of Housing and Urban-Rural Development proposed the first batch of pilot cities for the construction of CIM platforms. This marked the pivotal shift of China's CIM platform construction from theoretical exploration to practical implementation. In March 2021, the "14th Five-Year Plan" mentioned "improving urban information model platforms and operation management service platforms," pointing the direction for accelerating digital development and building a digital China. The "14th Five-Year Plans" of Shanghai, Jiangsu, and other provinces and cities have all mentioned CIM-related content, mainly focusing on the development of digital government, new infrastructure, and development of the construction industry. Consequently, the construction of CIM platforms has been fully rolled out, providing a more comprehensive urban scenario and policy support for the theory and application of CIM.

The research direction of CIM mainly includes urban governance, digital twin, intelligent construction site, intelligent environmental protection, etc. However, fewer studies have been carried out for the development of CIM construction, which is considered to be independent of the relationship with CIM research, and it is not possible to elucidate the interactive symbiotic relationship between CIM research and construction. This paper finds that CIM research guides the development of CIM construction by combing the relevant literature and CIM project bidding announcements in recent years, and CIM construction forces the progress of CIM technology, which promotes each other, advances the construction of a new smart city, and provides a better tool support for urban governance.

2 Research Methodology

2.1 Data Sources

The bibliometric analysis tool CiteSpace 6.1.R6 was utilized to comprehensively review and assess the overall situation and development trends of studies on urban governance and the relationship with CIM from 2008 to 2023. In constructing the search strategy, the subject was set to include papers with "City Information Model" in the subject terms. After excluding literature irrelevant to the research theme, a total of 414 Chinese literature were retrieved based on these subject terms.

Through filtering information from the China Tendering and Bidding Public Service Platform and various provincial and municipal government procurement websites and public service trading platforms, data collection, summary, and organization of CIM-related tender projects were completed. Based on this, the paper analyzes the current state of construction of CIM foundational platforms using the tender and bidding information database compiled from 2016 to July 2023.

2.2 Research Method

Bibliometrics enables a more objective evaluation of the development status of a discipline and more accurately identifies trends within the field. CiteSpace is an influential software for information visualization in the field of knowledge mapping, capable of intuitively presenting research hotspots and potential knowledge connections among literature in related fields [2]. This study utilizes CiteSpace to draw a visual knowledge map of CIM research, analyzing keywords, emerging terms, etc., and clarifying the current state of CIM research based on the analysis results.

Statistical analysis methods can concentrate and distill the information in the data in order to identify the inherent patterns in the object under study. This paper employs statistical analysis methods to organize and analyze data on CIM tender projects, providing a reference for a systematic understanding of the current state of CIM construction. Initially, Kernel Density Estimate (KDE) [3] is used to fit curves to the amounts of CIM tender projects, and the Kolmogorov-Smirnov test (K-S test) [4] is applied to evaluate the goodness of fit. Subsequently, polynomial regression is attempted to fit the curve of cumulative project numbers over time, and predictions are made based on the obtained polynomial. Finally, the Term Frequency-Inverse Document Frequency (TF-IDF) method [5] and the Latent Dirichlet Allocation (LDA) [6] method are utilized for keyword extraction and cluster analysis of tender project content, aiming to summarize or identify some patterns and issues in China's CIM construction.

3 Research Results

3.1 Analysis of the Current State of CIM Research

3.1.1 Annual Publication Volume and Trends.

The volume and trend of annual publications can generally reflect the importance and level of attention within the CIM research field. As illustrated in Figure 1, analysis of publications on CIM technology shows that scholarly output began in 2014. From 2014 to 2018, the growth was relatively slow, indicating that during this period, CIM-related research did not receive significant attention in China and was in its initial exploratory phase. After 2018, there was a noticeable increase in publications, indicating an increased focus within China on CIM-related research. However, the overall start of CIM-related research has been relatively recent, and the number of research outputs remains limited. However, the current volume of CIM research publications remains relatively low. There is a critical need to capitalize on the significant opportunities presented by the "14th Five-Year Plan" and the long-term goals for 2035 to accelerate research within China.

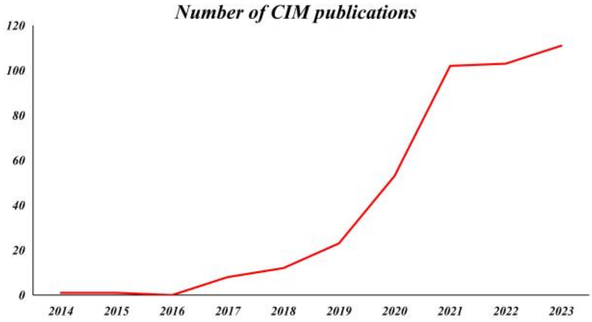


Fig. 1. Trend of publication volume

3.1.2 Keyword Clustering and Co-Occurrence Analysis.

The relational contributions of keywords are crucial for analyzing the research hotspots in a field. Multiple nodes connected by edges and clustered together, combined with metrics such as frequency and centrality, can visually present the important keywords, facilitating the identification of research hotspots in the field. Through keyword co-occurrence analysis, the overall research status and progression of CIM technology in the urban governance field from 2014 to 2023 were mastered. The keyword co-occurrence map contains 225 nodes and 353 links, with an overall network density of 0.014, as shown in Figure 2. Larger nodes and font sizes in the network indicate higher frequencies of keywords.

The modularity (Q) of the keyword clustering map is 0.617, and the silhouette coefficient (S) is 0.8984, indicating that the keyword clustering map is reasonable and credible. As shown in Figures 3, the map reveals that there are eight clusters. This demonstrates that the research hotspots in CIM technology for urban governance are quite extensive, with numerous application scenarios and significant potential for development.

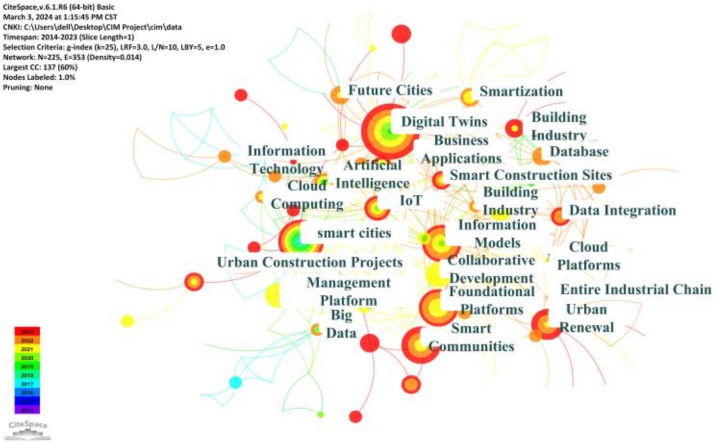
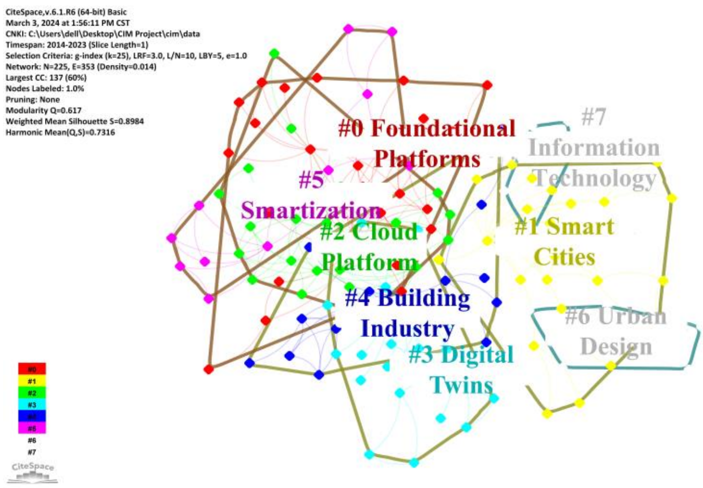


Fig. 2. Keyword co-occurrence network map



```
CiteSpace,v.6.1.R6 (64-bit) Basic
March 3, 2024 at 1:56:11 PM CST
CNKI: C:\Users\dell\Desktop\CIM Project\cim\data
Timespan: 2014-2023 (Slice Length=1)
Selection Criteria: g-index (k=25), LRF=3.0, L/N=10, LBY=5, e=1.0
Network: N=225, E=353 (Density=0.014)
Largest CC: 137 (60%)
Nodes Labeled: 1.0%
Pruning: None
Modularity Q=0.617
Weighted Mean Silhouette S=0.8984
Harmonic Mean(Q,S)=0.7316
```

Fig. 3. Keyword clustering map

3.1.3 Research Hotspots and Evolution.

The keyword burst map can display sudden increases or decreases in the frequency of citations within the literature, thereby reflecting significant shifts in research hotspots. To track pivotal changes in research hotspots within the urban governance domain related to CIM, the Burstness function of CiteSpace is used to analyze the sudden emergence of keywords. In terms of burst intensity, the research intensity of CIM is relatively weak, likely due to the recent emergence of CIM and fewer publications. Regarding the duration of bursts, the persistence of research hotspots in CIM studies is generally short, lasting less than three years.

Initially, the focal points of interest in CIM technology were centered on discrete foundational technologies such as big data and information technology. However, there has been a gradual shift towards the construction of application scenarios, such as the building industry, urban brains, and management platforms. The current focus has shifted towards the integration of applications and technologies, such as data integration and future cities.

3.2 Analysis of the Current State of CIM Construction

3.2.1 Statistical Analysis of Cim Project Tender Amounts.

In the distribution of tender amounts, KDE is used for curve fitting. As most projects are concentrated in lower monetary ranges, and a few projects have very high amounts, the distribution of tender amounts approximates a right-skewed distribution. Common right-skewed distributions include the log-normal distribution and the gamma distribution.

To identify the most effective fitting method, we used the K-S test to conduct a goodness-of-fit test on the data, as illustrated in Table 1.

Table 1. K-S test results

K-S test results	D-value	P-value
Log-normal distribution	0.0329	0.516
Gamma distribution	0.0998	Approximately 1.03×10^{-5}

In this test, a D-value of 0.0329<0.0998 indicates that the log-normal distribution has a relatively smaller maximum difference with the test sample compared to the gamma distribution. The P-value for the log-normal distribution is 0.516, which is much higher than the commonly used significance level (such as 0.05), indicating that the difference between it and the observed tender amount data is not significant, suggesting it is a reasonable model choice. On the contrary, the P-value is not a good fit.

According to the results of the K-S test, the log-normal distribution is more suitable as a model for tender amounts.

The curve of cumulative project numbers over time generally exhibits a nonlinear trend. Polynomial regression is used to attempt to fit this curve:

According to the fitting results, the polynomial model obtained is:

$$y = 29.99 - 24.80x + 3.37x^2 - 0.047x^3 \tag{1}$$

where y represents the cumulative number of projects and x stands for the number of quarters (counting from the first quarter of 2018).

The coefficients indicate that the growth in project numbers is not a simple linear relationship but is influenced by multiple factors, including the linear impact of time, acceleration effects, and more complex trends, as shown in Figure 4.

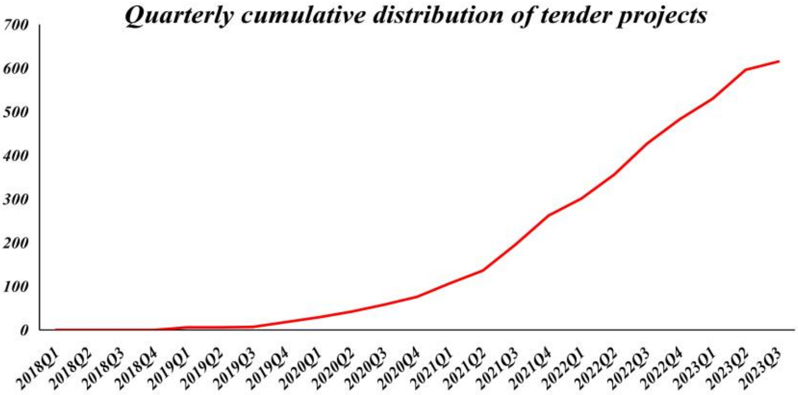


Fig. 4. Quarterly cumulative distribution of tender projects

Based on the polynomial model, predictions for the future are as follows:

The cumulative number of projects by 2025 is estimated to be approximately 947.58.

The cumulative number of projects by 2030 is estimated to be approximately 1421.09.

Despite a late start, the significant amount of investment suggests that, as time progresses, the cumulative number of projects is expected to continue to increase.

3.2.2 Clustering and Co-Occurrence Analysis of CIM Projects By Region.

TF-IDF is a common weighting technique used in information retrieval and data mining to assess the importance of a word to a document within a corpus.

LDA, as a commonly used topic modeling technique, can identify latent topics within a collection of documents.

3.2.2.1. National Analysis of CIM Construction.

(1) Prominence of smart cities and CIM technology: Nearly all themes emphasize terms such as "city," "data," "platform," and "CIM," indicating that the tender projects are generally related to the construction of smart cities, urban data management, and the application of CIM technology.

(2) Focus on technology and management: Keywords such as "platform," "system," "management," and "service" appear frequently, pointing to the importance of technology implementation and project management.

(3) Diversity of project characteristics: Although all themes share some common keywords, each theme has its unique focus, reflecting the diversity of the tender projects in terms of service content, technology application, and objectives.

3.2.2.2. Regional CIM Construction.

Based on the collected information on tender projects, an analysis of CIM construction in various regions is conducted, as illustrated in Figure 5.

By cleansing and processing the requirements and content of the projects, TF-IDF keywords are extracted (Table 2) to analyze the different construction focuses across regions.

4 Discussion of Results

Further analysis of the above research results indicates that China still faces several issues in the research and application of CIM:

From the perspective of CIM research, it is still in its early stages. The number of institutions and scholars researching CIM has gradually increased in recent years, and

the quantity of literature is also on the rise, indicating an increasing emphasis on CIM research. The main focus of CIM research is on the construction of CIM platforms and specific CIM scenario developments. However, there is a lack of systematic theoretical research on methods of CIM platform construction and key technologies. Although CIM research topics are quite broad and application scenarios are abundant, offering great developmental potential, there is an issue of breadth over depth. The research content often lacks detailed specificity, leading to a superficial exploration of complex topics.

From the perspective of the CIM application, several issues exist: The construction of CIM foundational platforms is regionally unbalanced, with rapid advancement in economically developed areas such as the southeast coast, while the construction in central and western regions is lagged. The focus of CIM+ applications is concentrated, and the application of CIM is not as comprehensive or widespread. Although substantial funds have been invested in the key construction of CIM foundational platforms, the overall development is still in its initial stages. During the construction of CIM foundational platforms, provinces and departments proceed independently, leading to a lack of unified data standards in system integration and the presence of data barriers. This fragmentation makes it difficult for platforms to achieve multi-departmental, multi-domain, and multi-disciplinary collaborative development in later stages.

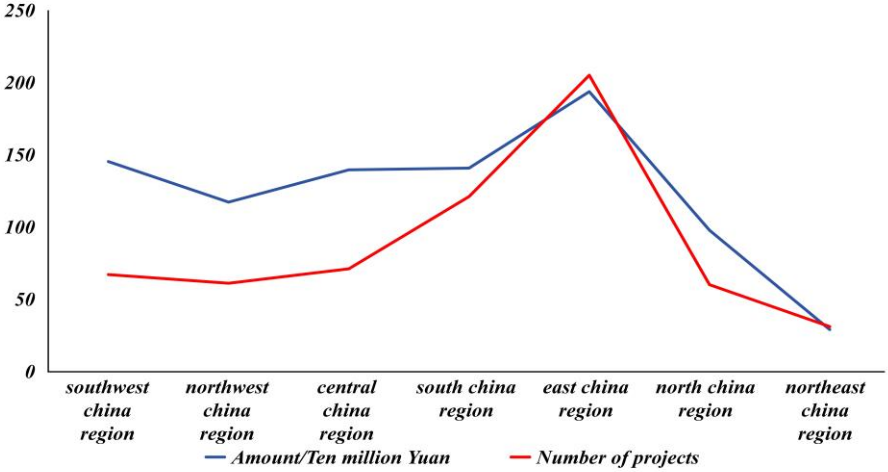


Fig. 5. Statistical analysis of tender project amounts and quantities by region

Table 2. Regional TF-IDF keywords

Region	TF-IDF keywords
North China	CIM, information, city, foundation, platform, data, smart, service, management, project
Northeast China	CIM, three-dimensional, information, underground, city, foundation, platform, data, model, management
East China	CIM, three-dimensional, information, city, foundation, platform, data, smart, model, management
South China	CIM, city, foundation, platform, data, service, model, management, system, project
Central China	CIM, information, city, foundation, platform, data, smart, service, management, system
Northwest China	City, platform, data, smart, service, monitoring, management, system, video, project
Southwest China	CIM, city, foundation, platform, data, smart, service, management, system, project

5 Conclusion

This paper has analyzed the basic information of recent CIM-related literature and construction projects to study the current state of research and construction of CIM. Both CIM-related research and construction are in their nascent stages and are experiencing rapid development. In terms of CIM research in China, interest began in 2014, with a rapid increase in activity from 2018, yet the overall volume of publications remains low. Research hotspots mainly focus on urban governance, smart cities, the IoT, and big data, with a wide range of application scenarios. In terms of CIM construction, the amounts of tender projects conform to a log-normal distribution, and predictions based on the cumulative tender amount curve suggest that the number of projects is expected to continue to grow. Although there are regional differences in CIM construction investment and content, the focus is consistently on building CIM foundational platforms. There are many issues in China's CIM construction that urgently need addressing.

To address the problems in CIM research and construction, at the theoretical and technical levels, the government should strengthen the top-level design and formulate development plans, while improving the standard system and promoting data interoperability. In terms of construction, the government should introduce financial subsidies and tax incentives to encourage enterprises and organizations to invest in the research, development and application of CIM, and at the same time promote data interoperability and sharing between different departments and systems to break down the information silos and realize the effective use of data, and finally select representative cities or regions to carry out demonstration projects of CIM construction, so as to provide lessons and references for other regions. This paper aims to offer some references for decision-making assistance and practical research in the field of CIM.

Acknowledgement

Funding: This study was supported by the National Key Research and Development Pro-gram (2022YFC3800600) and the National Natural Science Foundation of China for Young Scholars on the research of Implicit Knowledge Representation and Reuse Model in Design Review based on Immersive BIM and Case-based Reasoning (72001086).

Note: All figures and tables in the article are drawn by the author.

References

1. Chen Lei, Ji Jingling, Yan Xue. (2023) CIM+ Urban Operation Management Service Platform: Creating a New Model for Digital Urban Governance. Wisdom China, (Z1): 86-88. https://kns.cnki.net/kcms2/article/abstract?v=fsvnL9wA1q0LY8segxl-v7b2eBIW-peH4qPyWjnVLbxcThMzBVa63dM7jrQa6v6PTL-rhXPx72nv23ypcJQ6reuxzaOtIfzKDz6HQHoBdwH6_ustPe_B1WOF-DTXn0iqlKYO0Bt2NWyQ=&uniplatform=NZKPT&language=CHS.

2. Song Jianjun, Fan Chaoyu, Ren Ruowei, et al. (2024) Knowledge Map Analysis of Black and Odorous Water Body Research Based on CiteSpace. Express Water Resources & Hydropower Information: 1-13. http://kns.cnki.net/kcms/detail/42.1142.TV.20240417.1534.004.html.
3. Bryan S. Graham, Fengshi Niu, James L.Powell. (2024) Kernel density estimation for undirected dyadic data. Journal of Econometrics: Volume 240, Issue 2. 2024. PP 105336-.10.1016/J.JECONOM.2022.06.011.
4. Moscovich Amit. (2023) Fast calculation of p-values for one-sided Kolmogorov-Smirnov type statistics. Computational Statistics and Data Analysis. Volume 185,107769. https://doi.org/10.1016/j.csda.2023.107769.
5. Zhang Zhuo, Lei Yan, Mao Xiaoguang, et al. (2020) Fault Localization Approach Using Term Frequency and Inverse Document Frequency. Journal of Software: 31(11):3448-3460.DOI:10.13328/j.cnki.jos.006021.
6. Chen Erjing, Jiang Enbo. (2017) Review of Studies on Text Similarity Measures. Data Analysis and Knowledge Discovery, 1(06): 1-11. https://kns.cnki.net/kcms2/article/abstract?v=fsvnL9wA1q2da6gdLz4WIN0Ri5g-owHv_4ECLhZcIwToS9PucG4Y3IogTJAcQQdQochIZqg2c6q5E2S1W-Z6XBKbeY-JnSaFcKH3nq4fVRvQ-Wv-RIhahsJTpIM_TIsaBjkSI9tgn81U=&uniplatform=NZKPT&language=CHS.