CHARACTERIZATION OF THE CHINESE TRADITIONAL VILLAGES BASED ON THE MORPHOLOGICAL CLUSTERING AND KNOWLEDGE GRAPH

XIAO WANG[1], XUERONG ZHU[2] and PENG TANG[3]
[1,2,3]*School of Architecture, Southeast University, Nanjing 210096, China.*
[3]*Key Laboratory of Urban and Architectural Heritage Conservation (Southeast University), Ministry of Education, Nanjing, 210096, China.*
[1]*wangxiao_seu@seu.edu.cn, 0000-0001-9471-5370*
[2]*220220001@seu.edu.cn, 0009-0002-5640-885X*
[3]*tangpeng@seu.edu.cn, 0000-0003-1658-6774*

Abstract. The traditional settlements including Chinese traditional villages are facing the challenges posed by the digital divide in rural areas, which lack open-source data and comprehensive design. To address this, the study conducts an association analysis using morphological clustering and knowledge graphs, aiming to uncover the intrinsic logic and connections between tangible and intangible factors of village morphology. A comprehensive dataset of 8155 traditional villages, including geographical and morphological features, was compiled, supplemented with additional data on 1023 villages covering both tangible and intangible attributes. The methodology involves feature vector extraction using pre-trained neural networks, dimension reduction, and cluster analysis. Additionally, a graph database of village knowledge graph was established to identify entities and relationships, with visualization facilitated by the Neo4j. This research provides a method for analysing the characteristics of traditional villages, offering insights into their development trends and contributing to the formulation of globally applicable conservation and sustainable strategies in the era of artificial intelligence and climate change.

Keywords. Traditional Villages, Knowledge Graph, Clustering, Feature Vectors, Machine Learning.

1. Introduction

1.1. RESEARCH BACKGROUND

Traditional villages are complex self-organizing systems that have evolved through the ongoing interactions between human societies and their surrounding environments. This evolution has significant implications for ecological protection and sustainable development, especially today. As China experiences rapid development and

ACCELERATED DESIGN, Proceedings of the 29th International Conference of the Association for Computer-Aided Architectural Design Research in Asia (CAADRIA) 2024, Volume 2, 263-272.

urbanization, traditional villages encounter conflicts between economic growth and the conservation of cultural heritage. Modernization efforts have often led to a decline in the unique cultural characteristics and ecological integrity of many villages. Thus, recognizing and emphasizing the cultural value of traditional villages is crucial for their appropriate protection and inheritance in the process of modernization.

However, rural areas frequently suffer from a lack of digital infrastructure, positioning them on the fringes of the digital revolution. This scarcity of open-source data records and digital technology limits the adoption of advanced research that could potentially enhance the conservation of their traditional features. Therefore, it is essential to explore existing open-source data more thoroughly to overcome these challenges.

Beyond data-driven quantitative analyses, AI in specialized fields driven by knowledge is emerging and gradually being incorporated into architecture as a fundamental tool for decision-making. Machine learning (Tang et al., 2019), 3D digital models (Wang et al., 2017), and knowledge graphs (Cao, 2023) are widely used in urban texture research and preservation, offering valuable references. In particular, the potential of knowledge graphs in integrating and storing multi-dimensional associative data (Effendi et al., 2020) offers novel approaches for analysing large-scale (Yang and Qi, 2023) and high-precision morphological features (Yang et al., 2022).

For village feature research, nationwide studies on Chinese traditional villages primarily adopt a geographical perspective, viewing villages as points to analyse characteristics like distribution, equilibrium, and directional trends (Li et al., 2017). But most data-centric methods primarily focus on assessing and supporting design decisions related to morphology and tangible environments, while the intangible elements that fundamentally shape morphological properties are equally important.

To address this gap, we conducted an associative analysis using morphological clustering and knowledge graphs to explore the intrinsic logic and connections between the morphology of traditional Chinese villages and various intangible factors based on a case database. Clustering analysis methods, which identify latent structures and patterns in quantitative indicator datasets, enable researchers to understand the spatial differentiation characteristics of villages comprehensively and objectively. And by expanding the analytical perspective to include both tangible and intangible determinants, and enriching the scope of case studies, we propose a novel method for developing knowledge graph for information retrieval and visualization.

2. Methodology

2.1. DATA SOURCES AND COLLECTION METHODS

We constructed a dataset comprising 8155 traditional villages listed in official Chinese records, including geographic information like names, locations, and satellite and colored map images from open-source maps (Tianditu.gov.cn, 2023).

Due to the lack of metadata of each village, we collected related statistical data of 1023 villages that have been included from the "Traditional Chinese Village Digital Museum" (2023) website, which is still under construction. This data includes detailed tangible and intangible indicators, such as building area, population data, economic

indicators, and key industries. We employed one-hot encoding and dimension reduction for categorical variables, standardizing these data alongside continuous variables to eliminate the influence of dimensions or unites and magnitude on weighting. Subsequently, this data, combined with previously obtained morphological clustering types, was used for further analysis.

2.2. DATA ANALYSIS METHODS

To better analyse morphological clustering, we first extracted morphological elements and overlay them in one image. Then, we used a pre-trained neural network model CapsNet to extract feature vectors. For the required dimension reduction, we used the Uniform Manifold Approximation and Projection (UMAP) method, which efficiently reduces data complexity while preserving its global structure. After dimension reduction, we employed the Meanshift method for clustering analysis. By computing distances between feature vectors, Meanshift effectively groups villages according to their morphological similarity.

The entire analysis process was conducted in the Mathematica software environment. Mathematica's powerful computing capabilities and flexible data handling functions make it an ideal choice for conducting such complex data analyses. Through this methodology, we could more deeply understand the diversity and characteristics of traditional Chinese village forms.

In particular, we established a graph database after information extraction. The entity and relation are identified with the assistance of qualitative architectural ontology and social analysis, and attribute values are extracted from the crawled data. Using the Neo4j platform, the knowledge graph can be visualized and utilized according to the specified settings.

3. Clustering

3.1. MORPHOLOGICAL CLUSTERING OF VILLAGE OUTLINES

Building on our prior work in morphological clustering using satellite imagery, this study refines the process by enhancing feature extraction and clustering methods, and by employing advanced dimension reduction techniques for improved visualization. Although high-dimensional feature vectors extracted directly can be clustered, dimension reduction significantly decreases computational load and facilitates more intuitive visual interpretation of clustering outcomes.

The study synthesized road, water, and village outline data to create a simplified model encapsulating the spatial composition of both the built (buildings and roads) and natural (water) environments of villages (Figure 1). The feature exactor CapsNet, trained on the MNIST dataset for handwritten digit recognition, effectively captures the spatial hierarchy and feature relationships in image data, making it a suitable choice for processing overlay images with spatial composition features.

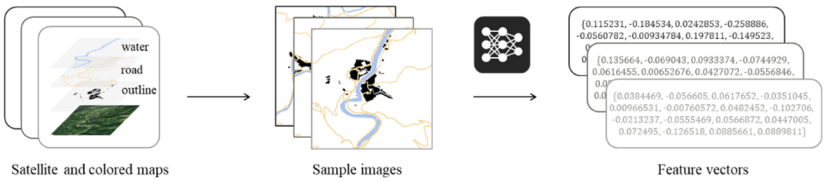

Figure 1. Image pre-processing and feature extraction

After obtaining feature vectors, we employed the UMAP algorithm for dimension reduction of high-dimensional feature vectors. This method maintains the global structure of data while effectively reducing its complexity. The spatial distribution of the two-dimensional feature vectors after reduction demonstrates a favourable clustering tendency. Using the MeanShift clustering algorithm, without pre-setting the number of clusters, we automatically obtained corresponding clustering results. As shown in the figures, the feature vector points are colored to distinguish clusters and their corresponding images are shown in the feature space (Figure 2). A diagram of all images arranged by similarity is also provided (Figure 3).

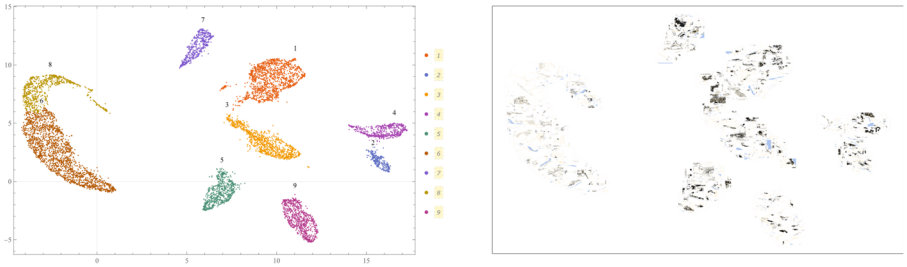

Figure 2. Morphological clustering results

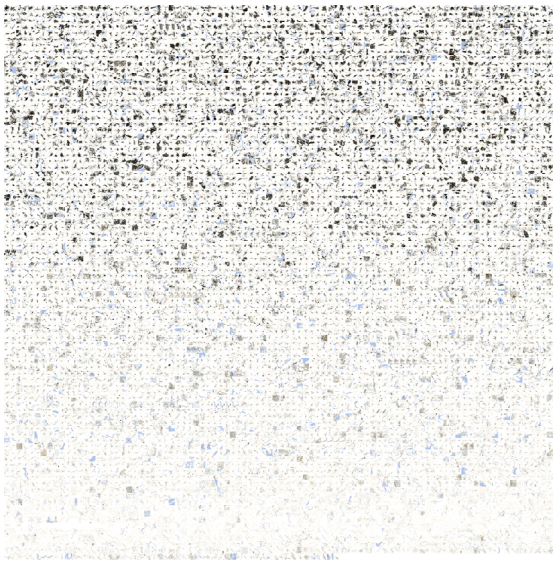

Figure 3. All samples sorted by similarity

3.2. CLUSTERING OF COMPREHENSIVE VILLAGE INDICATORS

Specifically, for the 1023 villages with socio-economic data obtained through web crawling, further analysis was conducted. The accessible statistical indicators cover three aspects: built environment, geographical environment, and social environment. The built environment includes parameters like village administrative area and building area; the geographical environment encompasses altitude and main terrain; the social environment includes formation era, registered population, resident population, total annual income, per capita annual income, main ethnicity, and key industries. Indicators like main terrain, formation era, main ethnicity, and key industries, comprising several independent types, are categorical variables. Others, such as altitude and area, are continuous variables. We applied one-hot encoding and dimension reduction to transform categorical variables into numerical values and standardized these along with other continuous variable indicators to form normalized feature vectors. The corresponding comprehensive indicator clusters (Figure 4) as follows:

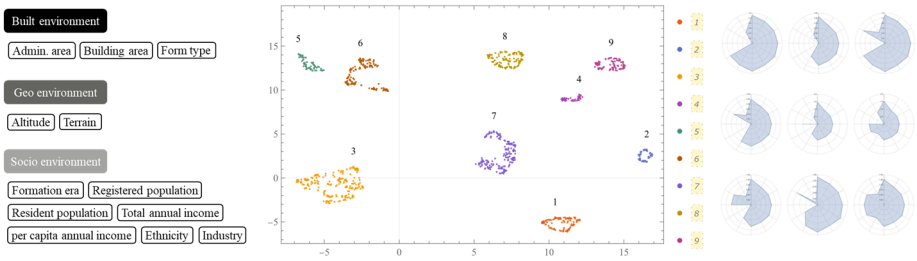

Figure 4. Indicators and clustering results

3.3. EVALUATION OF CLUSTERING RESULT

The Silhouette Coefficient (SC) is used to evaluate clustering effectiveness, reflecting both the cohesion and separation of clusters. It is a common assessment method with values ranging from -1 to 1, where a higher coefficient indicates better clustering (Yang et al., 2023). The silhouette coefficients for overall and each type in morphological clustering and indicator clustering are calculated as follows (Figure 5):

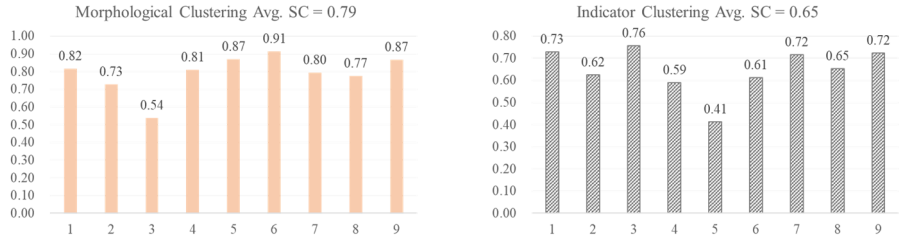

Figure 5. SC for overall and each type in morphological and indicator clustering

For the nine categorized village morphologies, representative village samples were identified at the clustering centres. While it is possible to describe the morphological characteristics of each sample as in previous studies, our experiment (which potentially

could get 6 to 11 different clusters using other methods), found it more interesting to observe the various stages of village morphological change and the fusion forms that arise from different environmental variations (Figure 6).

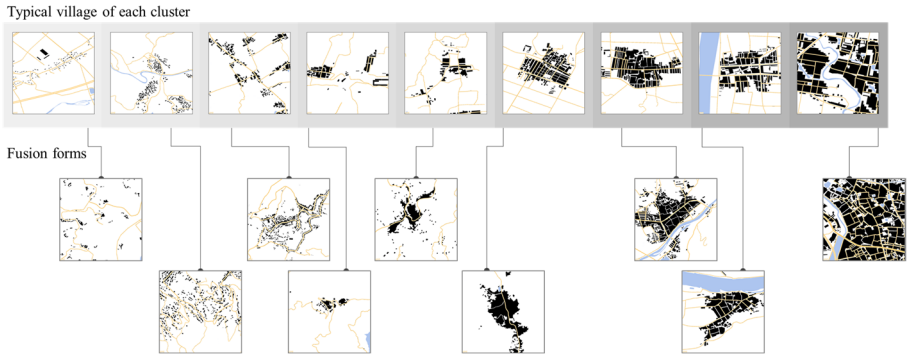

Figure 6. Typical and fusion forms of village clusters

We annotated samples by locations to obtain a spatial distribution across mainland China (Figure 7). Unlike previous studies (Wang et al., 2023), the distribution of types based on a broader sample set (both in terms of location and density) suggests that factors influencing village morphology may not be large-scale geographical environment, nor administrative divisions conventionally considered. They are more likely to be medium-scale and micro-scale climatic, terrain and humanistic factors.

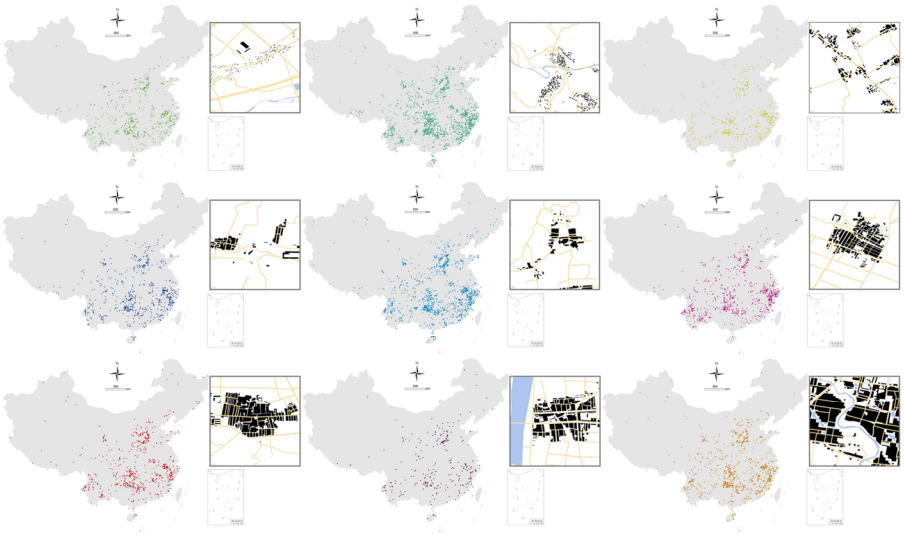

Figure 7. Geographical distribution of village clusters

For the clustering results of comprehensive indicators, we obtained characteristic features for each type. The similarity of feature vectors derived from clustering will serve as the foundation for subsequent knowledge graph construction.

4. Knowledge Graph

The Knowledge Graph, a key branch of AI, was introduced by Google in 2012. It is a structured semantic knowledge base that symbolically describes the concepts in the physical world and their interrelationships. The fundamental unit is the "entity-relationship-entity" triplet, along with entities and their associated attribute-value pairs. These entities are interconnected through relationships, forming a net-like knowledge structure (Johnson et al., 2022). Integrating data collected and cleaned from the previous stages and feature vectors from clustering analysis as data sources, and combining domain knowledge with graph structures, we constructed a comprehensive knowledge graph of Chinese traditional villages on the Neo4j platform (Figure 8).

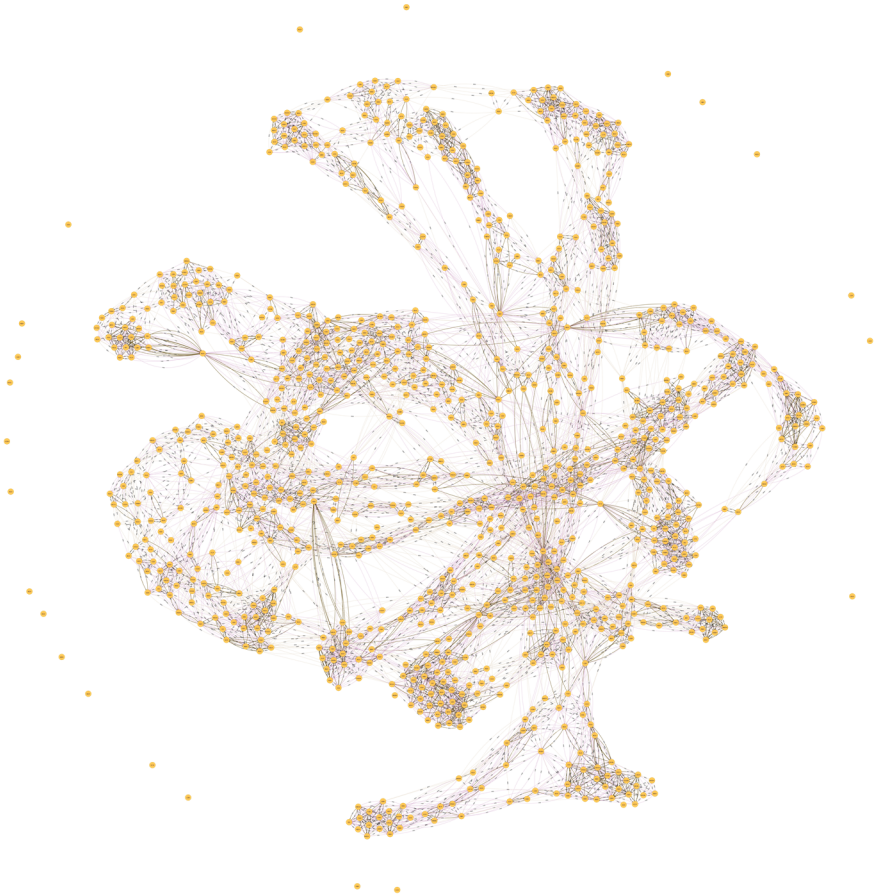

Figure 8. Knowledge graph of Chinese traditional villages

4.1. ENTITIES RECOGNITION AND ATTRIBUTES

Taking the geographical, social, and built environments of villages as primary cognitive variables, and traditional villages as the main information entities, we established a multi-level qualitative representation framework.

This framework allows entities to be described through cognitive variables. The first level consists of attribute indicators, such as terrain, total annual income and village area. The second level includes descriptive indicators, further subdividing the category indicators to differentiate the characteristics of villages, such as high altitude, low per capita annual income. The third level comprises village entities, which are the main objects for storing information and establishing relationships. First, we connect the first and second level indicators to form a feature attribute index system (Figure 9). Then, using the third-level village entities as nodes and village environmental attributes as edges, we connect village groups with descriptive indicators, completing the pathway for querying village characteristics.

4.2. DEFINING RELATIONSHIPS BETWEEN ENTITIES

The similarity relationships between villages are determined based on feature vectors calculated from clustering analysis, selecting the 15 villages with the closest feature vectors as a similar village group. We use 30% and 70% thresholds of the Euclidean distance of all similarity village feature vectors as criteria to define strong, medium, and weak similarity relationships between villages (Figure 9).

Similar features are those with the same category and similar values. Based on these criteria, we establish similarity relationships between related village nodes and store the similarity features in the edge attributes, fully expressing the interconnections among village groups.

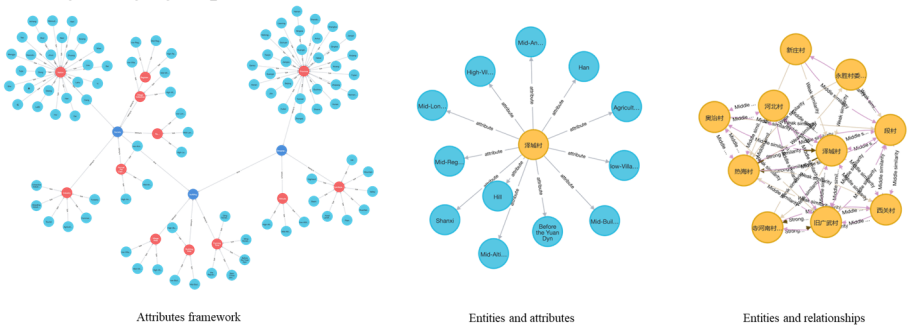

Figure 9. Entities, attributes and relationships of knowledge graph

4.3. CONSTRUCTING AND UTILIZING THE KNOWLEDGE GRAPH

The Knowledge Graph organizes extensive of web data in a structured and searchable format, supported by a "search + knowledge base" approach. It enables efficient querying of entities and their relationships via the Cypher query language in Neo4j, presenting information in a user-friendly visual format.

On the one hand, the qualitative representation framework's connections allow for a quick understanding of basic village features, and through filtering feature information, villages meeting specific criteria can be identified (Figure 10). On the other hand, based on the similarity relationships between village entities, similar village groups can be queried, providing case references for the self-renewal of individual villages (Figure 11). For instance, based on existing indicators, for searched villages with similar natural environments such as altitude and terrain compared with research case, when the village area is also similar, the villages with higher income status can serve as references for renewal development in terms of proper population composition and industry.

Figure 10. Searching by specific attribute

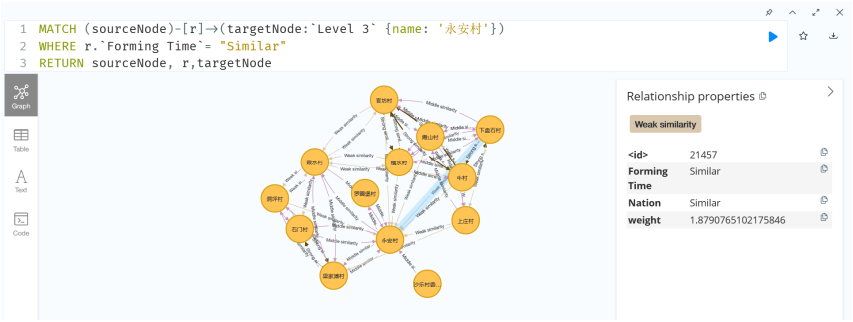

Figure 11. Searching by specific relationship

5. Conclusion

This research conducted a comprehensive and precise analysis of villages using both villages outline images and socio-economic statistical indicators:

Through improved image clustering, different representative types of village morphology were identified, revealing the developmental trajectory of Chinese traditional villages. By clarifying the relationship between the tangible and intangible elements and constructing a knowledge graph, we achieved structured representation of traditional village information.

In this work, morphological clustering provides an effective means for understand of the spatial structure and morphological features, deepening our understanding of village development patterns. Utilizing knowledge graphs to integrate and correlate data on built, natural and social environments of villages reveals their multi-dimensional development characteristics, offering a scientific basis for village

conservation and sustainable development.

In the future research, more text and images from social media and historical documents can be considered to enrich the information in knowledge graph. It would undoubtedly deepen the contextual understanding of traditional villages, facilitating a more nuanced analysis of their evolution and current state.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 52178008, Postgraduate Research & Practice Innovation Program of Jiangsu Province under Grant KYCX21_0110 and China Scholarship Council (No. 202206090272).

References

Cao, X. (2023). *Construction and representation of urban historical and cultural landscape assessment model based on knowledge graph: a case study of Beijing.* Beijing University of Civil Engineering & Architecture.

Effendi, S. B., van der Merwe, B., & Balke, W. T. (2020). Suitability of Graph Database Technology for the Analysis of Spatio-Temporal Data. *Future Internet,* 12(5), 1-31.

Johnson, J. M., Narock, T., Singh‐Mohudpur, J., Fils, D., Clarke, K. C., Saksena, S., Shepherd, A., Arumugam, S., & Yeghiazarian, L. (2022). Knowledge graphs to support real‐time flood impact evaluation. *AI Magazine,* 43(1), 40-45.

Li, Y., Yao, W., Zhang, Y., & Peng, W. (2020). Spatial distribution characteristics of Chinese traditional villages. *China Cultural Heritage,* (04), 51-59.

National Platform for Common Geospatial Information Services. Retrieved December 16, 2023, from https://www.tianditu.gov.cn/

Tang, P., Li, H., Wang, X., & Hovestadt, L. (2019). Generative design on conservation and inheritance of traditional architecture and settlement based on machine learning: a case study on the urban renewal design of Roma termini railway station. *The Architect,* (01), 100-105.

Traditional Chinese Village Digital Museum. Retrieved December 16, 2023, from http://www.dmctv.cn/

Wang, L., Fang, K., Xie, H., & Xiong, W. (2017). 3D urban design platform construction and innovation. *Planners,* 33(02), 48-53.

Wang, X., Tang, P., & Cai, C. (2023). Traditional Chinese village morphological feature extraction and cluster analysis based on multi-source data and machine learning. In *28th International Conference on Computer-Aided Architectural Design Research in Asia: Intelligent and Informed, CAADRIA 2023* (pp. 179-188). The Association for Computer-Aided Architectural Design Research in Asia (CAADRIA).

Wolfram Mathematica. Retrieved December 16, 2023, from https://www.wolfram.com/mathematica/

Yang J., Shao, D., Wang, P., Yin, S., & Murong, Z. (2022). Integration, topology, and translation: an in-depth analysis method of urban form based on knowledge map. *City Planning Review,* 47(06), 57-67.

Yang, C., & Qi, G. (2023). An Urban Traffic Knowledge Graph-Driven Spatial-Temporal Graph Convolutional Network for Traffic Flow Prediction. In *11th International Joint Conference on Knowledge Graphs, IJCKG 2022* (pp. 110-114).

Yang, Z., Yang, W., & Li, J. (2023). Research on the classification of residential buildings in urban blocks based on k-means. *Urban Environment Design,* 143(06), 360-364.