Integrating Point of Interest and BERT to identify potentially contaminating Enterprises in Datong City

Xin Du[1], Bin Yang[1], Jing Tong[2], Xiaomin Du[2], Liya Liu[1], Jinzhong Yang[2], Weiling Yao[2]

[1] Hebei Remote Sensing Center
[2] China Aero Geophysical Survey and Remote Sensing Center for Natural Resources

Keywords: Contaminated enterprises; POI, BERT; Datong city; Polluting industries

Abstract

Effective management of environmental safety risks in brownfield redevelopment relies on accurate identification of contaminated enterprises. A key challenge is the rapid acquisition of data on these enterprises. This study proposes a method leveraging Point of Interest (POI) data and a BERT-based prediction model to identify potentially contaminated enterprises. The method was applied to Datong, a major industrial and mining city in China. The method successfully identified 329 potentially contaminated enterprises across 23 types of polluting industries. Notably, enterprises in the mining and washing sectors of bituminous and anthracite coal represented 26.2% of the total identified, reflecting Datong's coal-centric industrial nature. The proposed method efficiently identifies potentially contaminated enterprises, supporting targeted environmental risk management and brownfield redevelopment. Integrating it with regulatory frameworks can enhance compliance monitoring and inform decision-making for sustainable urban development.

1. Introduction

In recent years, with the rapid urbanization and the growing scarcity of construction land in China, land formerly occupied by closed industrial and mining enterprises has gradually become a critical target of urban redevelopment and regeneration initiatives(Wu et al., 2024). However, However, a substantial proportion of these industrial and mining lands are potentially contaminated, and many pollution-related activities remain insufficiently investigated or are still at an early stage of assessment(Huang et al., 2020). Historical industrial production and mining activities conducted on these sites have introduced persistent environmental pollutants and latent ecological risks, which may threaten public health and residential safety during subsequent land redevelopment and reuse processes. Consequently, the effective identification and management of potentially contaminated enterprises have become significant challenges for local governments in the context of sustainable urban governance. At present, most local government agencies possess only limited statistical records of industrial enterprises, while accurate geographic location information is often incomplete or unavailable(Li et al., 2019, Li et al., 2022). Therefore, accurately identifying the quantity, industrial characteristics, and spatial distribution of potentially contaminated enterprises has become an urgent issue for urban managers seeking to implement stricter land-use planning regulations and environmental governance policies.

Previous studies have demonstrated that although a large number of potentially contaminated sites exist worldwide, only a relatively small proportion has been effectively identified and systematically documented(Mostafavi et al., 2021). To address this issue, some researchers have relied on publicly available governmental databases and authoritative institutional records to identify polluting enterprises, subsequently verifying the acquired information through statistical and tabular analysis methods(Qiu et al., 2023, Sahu et al., 2024). Although these statistical approaches can rapidly generate preliminary identification results, their accuracy is highly dependent on data completeness, reliability, and timeliness, and they often require integration with supplementary analytical approaches for comprehensive evaluation. In recent years, advances in Natural Language Processing (NLP) technologies have provided new opportunities for the large-scale acquisition and identification of contaminated enterprise information(Li et al., 2018, Li and Gu, 2024, Ma et al., 2023). With the increasing maturity of NLP applications in text mining, semantic analysis, machine translation, and opinion mining, researchers have begun to employ Google Point of Interest (POI) data to predict enterprise industrial categories and identify potentially contaminating enterprises(Huang et al., 2023). Nevertheless, existing studies mainly focus on the identification of a single category of contaminated enterprise, thereby limiting the generalizability and practical applicability of these methods in complex urban industrial environments.

In this study, we integrate POI (Points of Interest) data with BERT (Bidirectional Encoder Representations from Transformers)(Devlin et al., 2019) to construct a comprehensive identification framework2 for the systematic and accurate identification of potentially contaminated industrial enterprises in Datong City, a major industrial and coal-mining city in China. First, enterprise POI data were collected from the Gaode Map open platform through web crawler technology, and multi-source data fusion methods were subsequently employed to construct a dataset containing approximately 4,000 enterprise records in Datong City. Second, based on Google's open-source BERT pre-trained model, we fine-tuned the model using pollution permit records and enterprise business registration data from Shanxi Province to develop a BERT-based classification model capable of predicting potentially contaminated enterprise categories. Finally, the trained BERT-based prediction model was applied to the large-scale POI dataset of Datong City to identify contaminated industrial categories and screen potentially polluting enterprises. Overall, this study provides a scalable and data-driven methodological framework for the rapid identification and spatial management of potentially contaminated enterprises, which may support urban environmental governance and sustainable land-use planning in resource-based cities.

2. Materials

2.1 Study areas

The study area covered in this article spans latitudes $39°3'$ to $40°44'$ and longitudes $112°34'$ to $114°33'$, located in the northern Shanxi Province, fig1. Datong City is one of China's largest coal energy bases, renowned for its coal mining and mineral washing and separation activities. With a long industrial history, Datong has developed a series of energy chemical industries based on its coal resources, including coal-to-gas, coal-to-oil, coal-to-chemical fertilizer, and other related industrial enterprises. This study focuses on identifying the contaminated enterprises in Datong that are prone to causing environmental pollution.

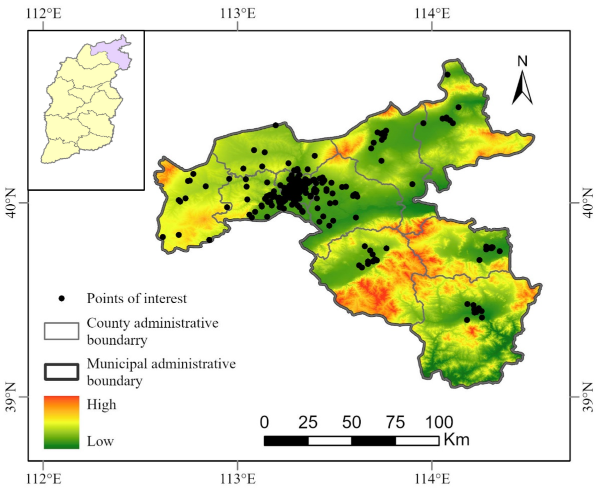

Figure 1. Spatial Distribution Map of Enterprises in Datong City.

2.2 Data sources

The POI data for Datong's enterprises, including latitude and longitude locations but lacking specific enterprise properties, were acquired from the point data services of Gaode's open-source electronic map (https://www.amap.com). The semantic information related to the enterprises, such as names and business scopes, was partially obtained from the national emission permit management platform's database of enterprise emission permits (http://permit.mee.gov.cn/permitExt), and further supplemented by data from enterprise search websites and Bing Search. The POI data and the acquired semantic information were then processed through data fusion to create a foundational dataset of suspected polluting enterprises. The derived enterprise properties data were used to train and optimize the BERT prediction model.

3. Methodology

3.1 Overview of the methodology

This paper proposes a novel method for predicting the distribution of potentially contaminated enterprises using massive multi-source big data and automated computer processing technology, eliminating the need for field investigations of contaminated sites. The method consists of three key components, fig2. The first component involves acquiring enterprise POI data, where web crawling techniques are used to collect POI data for Datong City, initially with low property accuracy. This is followed by retrieving and obtaining enterprise names and business scopes based on the POI names, ultimately constructing a dataset of Datong enterprise POIs that includes business scopes. The second component focuses on training a BERT prediction model for enterprise business classification. During the BERT pre-training phase, the model learns features of natural language text, which are then utilized in the fine-tuning phase for specific tasks. The enterprise industry classification model is fine-tuned using the enterprise business scopes as input text and industry categories as labels. Finally, after the model is validated, it is applied to predict the industry categories of the Datong enterprise POI dataset, identifying suspected polluting enterprises within polluting industries. This method not only enables the identification of potentially contaminated enterprises but also provides insights into their spatial distribution.

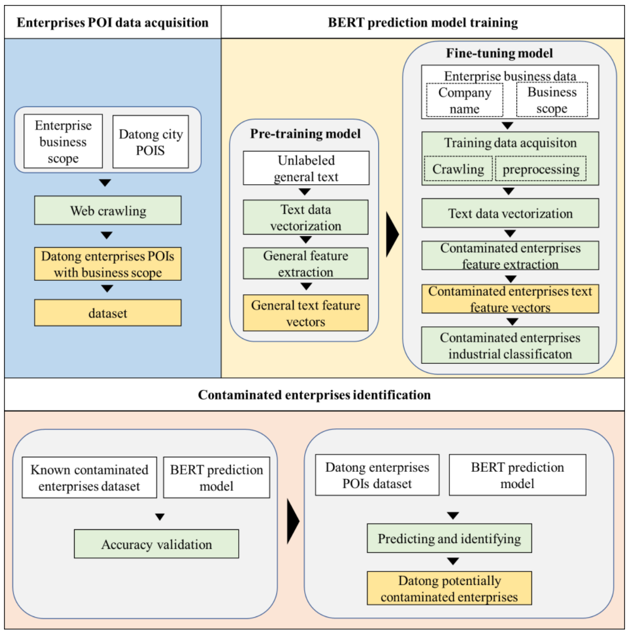

Figure 2. Methodology flowchart.

3.2 POI data acquisition for enterprises

POI (Point of Interest) data, referring to geospatial point data associated with specific urban entities and functions, is typically acquired through web crawling techniques as an important component of digital map services and has been widely applied in urban spatial analysis and geographic research(Huang et al., 2022, Liu et al., 2019, Tu et al., 2022, Yang et al., 2024).

In this study, a web crawler approach combined with the quadratic partitioning of administrative districts was employed to systematically acquire comprehensive enterprise POI data for Datong City from the Gaode Map open platform. Specifically, Python scripts were developed to send HTTP GET requests to the Gaode online map server, receive the returned response data, and subsequently parse the relevant information to extract enterprise-related POI records. The extracted POI dataset primarily included enterprise names, geographic coordinates, address information, and category attributes, thereby providing an essential spatial database for subsequent analysis.

Among these variables, latitude and longitude information constituted critical spatial parameters for the subsequent BERT-based identification model used to recognize potentially contaminated enterprises. To ensure the spatial accuracy and reliability of the acquired coordinates, spatial matching and administrative district verification procedures were further conducted on the POI data. In addition, data related to sewage discharge enterprises, enterprise registration records, and fragmented enterprise information in Datong City were collected from official governmental websites using Python-based automated data extraction methods. The business scope field extracted from enterprise registration information was subsequently integrated with the Datong enterprise POI dataset to supplement semantic and industrial attribute information, thereby providing critical foundational data for the BERT model to predict and identify potentially contaminated enterprises.

After data cleaning, spatial verification, and multi-source information integration, a total of 4,022 enterprise POI records with accurate geographic locations and complete business scope information were obtained for Datong City, as presented in Table1.

3.3 BERT-based prediction model training

BERT, short for Bidirectional Encoder Representations from Transformers, is a pre-trained sequence-based language model released by Google in 2018, which utilizes an Embedding layer and multiple Transformer layers to extract the features of the input text, and finally outputs vectors containing text features, can be used for different natural language processing tasks(Devlin et al., 2019, Li et al., 2020, Kaliyar et al., 2021, Brandes et al., 2022, Acheampong et al., 2021). In this study, we introduced a novel BERT-based prediction model for identifying potentially contaminated enterprises, which comprises both model pre-training and fine-tuning.

The pre-training phase of BERT serves as the process of initializing the model parameters. These pre-trained parameters capture rich linguistic patterns, which significantly reduces the training time for subsequent models and enhances their performance. In this study, we used the BERT-Base-Chinese model released by Google, which is pre-trained on a corpus from Chinese Wikipedia containing 25 million sentences. This model is composed of a 12-layer Transformer encoder with 768 hidden units and 110 million parameters.

Fine-tuning adapts the pre-trained model to the specific task of identifying potentially contaminated enterprises. A labeled dataset was constructed using industry category fields from actual contaminated enterprises in Shanxi Province, based on pollution permit data and business records. This dataset includes 23 contaminated industry categories, such as brick and stone manufacturing, metal casting, and the production of cement, lime, and gypsum.

During fine-tuning, the labeled dataset was used to update the model parameters to improve classification performance. An additional fully connected layer and a Softmax layer were added to the top of the BERT model for text classification. The input to the fully connected layer is the hidden representation from the final Transformer layer. This layer consists of multiple neurons, fully connected to the previous layer, with its structure adjustable based on task complexity and dataset characteristics.

To enhance model expressiveness, a Tanh nonlinear activation function was applied in the fully connected layer. The Softmax function in the output layer converts the model's output into probability distributions across label categories, enabling effective classification. The fine-tuning parameters were optimized through iterative adjustments and validations, as detailed in Table 2.

3.4 Contaminated enterprises identification and validation

This study employed a BERT prediction model—both pretrained and fine-tuned—to classify potentially polluting enterprises within the business industry POI dataset of Datong City. The model used the Softmax function to assign the category with the highest probability as the final classification result. A total of 329 potentially polluting enterprises were identified in Shanxi Province, covering 22 sub-industries, as shown in Table 3.

To evaluate the model's classification accuracy, the cross-entropy loss function was applied. This assessment compared the number of correctly identified enterprises within each industry category in the test dataset to the total number of enterprises in the test set. Additionally, model performance was validated using a list of 22 key polluting enterprises published by the Datong City Ecological Environment Bureau. The model successfully identified 16 out of the 22 enterprises, achieving an actual recognition accuracy of 73%, which satisfies the accuracy requirements for large-scale data screening. Selected validation results are presented in Table 4.

4. Results

4.1 Distribution of the potentially contaminated enterprises in Datong

Kernel Density Estimation (KDE) is a statistical technique used to estimate the probability density function of a random variable. In this study, KDE was employed to analyze the spatial distribution of potentially contaminating enterprises within the study area. The results reveal that density hotspots are primarily located in the mid-northern region. On the map, varying density values are represented by different colors, with darker shades indicating higher concentrations of potentially contaminating enterprises.

4.2 Performance of the BERT training models

The method proposed in this study offers a promising approach for identifying potentially contaminated enterprises. By integrating POI data with a BERT-based predictive model, it facilitates rapid identification, significantly reducing the time and resources required compared to traditional site-by-site assessments. This approach successfully identified 329 potentially contaminated enterprises across 22 polluting industry categories, providing a comprehensive tool for environmental risk management. Notably, the identification of enterprises involved in the mining and washing of bituminous and anthracite coal aligns with Datong's coal-dominated industrial profile.

However, several limitations must be acknowledged to ensure the robustness and generalizability of this methodology. The

ID	Province	City	POI Name	Longitude	Latitude	POI Category	Alias
1	Shanxi Province	Datong City	Hanfengling Wind Farm	113.93093	39.63136	Company	Other
2	Shanxi Province	Datong City	Gaoyuanhong Agriculture Co., Ltd.	113.96545	39.79769	Company	Company
3	Shanxi Province	Datong City	Yuhong Breeding Farm	113.90127	39.95903	Company	Agriculture, Forestry, Animal Husbandry, and Fishery
4	Shanxi Province	Datong City	Shuntian Group Huanghua Base	113.94688	39.96977	Company	Agriculture, Forestry, Animal Husbandry, and Fishery
6	Shanxi Province	Datong City	Guiyi Carbon Co., Ltd.	113.46954	39.88406	Company	Company

Table 1. Selected POIs in Datong City.

Parameter Name	Parameter Value	Parameter Meaning
task_name	comp	Training Task Name
do_train	true	Whether to Train
do_eval	true	Whether to Evaluate Accuracy
data_dir	../MyData/Company predict	Data Storage Directory
vocab_file	chinese_L-12_H-768_A-12/vocab.txt	Vocabulary File Path
bert_config_file	chinese_L-12_H-768_A-12/bert_config.json	BERT Configuration File Path
init_checkpoint	BERT_BASE_DIR/bert model.ckpt	Initial Checkpoint Path
max_seq_length	100	Maximum Text Sequence Length

Table 2. Fine-tuning Task Parameters.

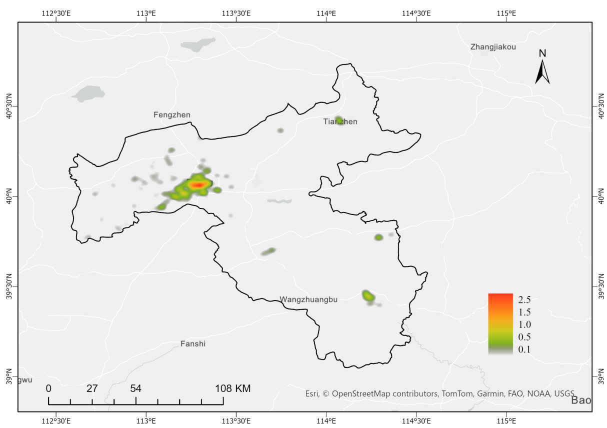

Figure 3. Kernel Density Map of Polluting Enterprises in Datong City.

accuracy of the identification process is highly dependent on the quality and completeness of the POI data. Outdated or inaccurate data may lead to misidentification. Moreover, the reliance on industry-specific information is constrained by the availability and reliability of such data. Information obtained from enterprise search websites and Bing Search may be outdated or inaccurate, failing to reflect current enterprise activities or

Industry Category	Total
Electricity production	5
Seasoning and fermentation product manufacturing	10
Non-metallic waste and debris processing	4
Fertilizer manufacturing	4
Environmental management industry	19
Basic chemical raw material manufacturing	8
Alcohol production	5
Wholesale of mineral products, building materials, and chemical products	6
Refractory material products manufacturing	3
Sales of automobiles, motorcycles, parts and accessories, fuel, and other types of power	8
Production and supply of heat energy	4
Manufacture of graphite and other non-metallic mineral products	27
Manufacture of cement, lime, and gypsum	6
Manufacture of ceramic products	3
Manufacture of paint, ink, pigments, and similar products	6
Slaughtering and meat processing	3
Mining and preparation of bituminous and anthracite coal	81
Hospitals	10
Manufacture of paper products	15
Manufacture of castings and other metal products	26
Manufacture of specialized chemical products	11
Manufacture of bricks, tiles, stone, and other building materials	65

Table 3. Identification Results of Potential Polluting Enterprises in Datong City Based on the BERT Prediction Model

emissions. The integration of multiple data sources, including the national emission permit management platform, may also result in incomplete coverage, as some enterprises may not be included or may underreport their emissions.

5. Conclusions

The method proposed in this study, which integrates POI data with BERT-based predictive models, demonstrates considerable potential for enhancing environmental risk assessment and management practices. By enabling the rapid identification of potentially contaminated enterprises, the proposed approach supports more targeted and efficient intervention strategies, particularly in the context of brownfield redevelopment. Early identification of such sites may contribute to the prevention or mitigation of adverse public health impacts and environmental degradation.

Number	Company Name	Identified or Not
1	Shanxi Xinguang Industry and Trade Co., Ltd	Yes
2	Jinneng Holding Group Jinyin Mining Co., Ltd	Yes
3	Shanxi Jindi Mining Co., Ltd	Yes
4	Lingqiu County Yiheng Mineral Processing Plant Branch 1	No
5	Lingqiu County Tianshi Trading Co., Ltd	No
6	Lingqiu County Hengtai Non-Ferrous Metals Smelting Co., Ltd	Yes
7	Lingqiu County Fushen Mining Products Development Co., Ltd	Yes
8	Shanxi Donghua Machinery and Electronics Co., Ltd	Yes
9	Shanxi Nairui Synthetic Rubber Co., Ltd	Yes
10	Shanxi Tianding Juyuan Silver Industry Co., Ltd	No
11	Shanxi Shigong Cement Co., Ltd	Yes

Table 4. Verification of Identified Suspected Polluting Enterprises (Partial)

Furthermore, integrating the proposed framework with existing regulatory systems and databases, such as the national emission permit management platform, could further enhance its practical applicability for environmental governance. Alignment with these platforms may improve the accuracy and timeliness of compliance monitoring, facilitate the identification of non-compliant enterprises, and strengthen the tracking and management of pollutant emissions. Consequently, the proposed method could support more effective enforcement of environmental regulations and regulatory oversight.

As a decision-support tool, the proposed method can assist in informing zoning policies, prioritizing site remediation efforts, and optimizing the allocation of environmental protection resources. By providing policymakers and urban planners with timely and reliable information, the framework can support sustainable urban development and promote more effective environmental stewardship.

Founding: This research was funded by Xiong’an New Area Special Program for Science and Technology Innovation (2024XAGG0015): Development of Key Technologies for the Intensive Utilization and Sustainable Value Enhancement of Territorial Space,

References

Acheampong, F. A., Nunoo-Mensah, H., Chen, W., 2021. Transformer models for text-based emotion detection: a review of BERT-based approaches. *ARTIFICIAL INTELLIGENCE REVIEW*, 54(8), 5789–5829.

Brandes, N., Ofer, D., Peleg, Y., Rappoport, N., Linial, M., 2022. ProteinBERT: a universal deep-learning model of protein sequence and function. *BIOINFORMATICS*, 38(8), 2102–2110.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. J. Burstein, C. Doran, T. Solorio (eds), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186.

Huang, C., Xiao, C., Rong, L., 2022. Integrating Point-of-Interest Density and Spatial Heterogeneity to Identify Urban Functional Areas. *REMOTE SENSING*, 14(17).

Huang, G., Zhu, S., Wang, X., Tian, Z., Ji, G., Lu, R., Cui, X., Chen, X., 2020. Natural language processing and machine learning-based suspected soil contamination enterprise identification. *Chinese Journal of Environmental Engineering*, 14(11), 3234–3242.

Huang, S., Wu, X., Wu, X., Wang, K., 2023. Sentiment analysis algorithm using contrastive learning and adversarial training for POI recommendation. *Social network analysis and mining*, 13(1), 75–75.

Kaliyar, R. K., Goswami, A., Narang, P., 2021. FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *MULTIMEDIA TOOLS AND APPLICATIONS*, 80(8), 11765–11788.

Li, C., Sun, T., Zhai, L., Yuan, K., 2019. City Profile: Datong, China. *Environment and Urbanization ASIA*, 10(2), 176–192. https://doi.org/10.1177/0975425319863928.

Li, F., Gui, Z., Wu, H., Gong, J., Wang, Y., Tian, S., Zhang, J., 2018. Big enterprise registration data imputation: Supporting spatiotemporal analysis of industries in China. *COMPUTERS ENVIRONMENT AND URBAN SYSTEMS*, 70, 9–23.

Li, R., Gu, B., 2024. Enterprise ecological niche measurement and competitive effect analysis based on digital business recognition. *TECHNOLOGY ANALYSIS & STRATEGIC MANAGEMENT*, 36(11), 3621–3636.

Li, X., Liu, Z., Wu, K., Lu, Q., Shen, M., Zhou, Y., 2022. Analysis of Urban Land Development Density with POI Modeling. *Journal of Geomatics*, 47(5), 98–101.

Li, X., Zhang, H., Zhou, X., 2020. Chinese clinical named entity recognition with variant neural structures based on BERT methods. *JOURNAL OF BIOMEDICAL INFORMATICS*, 107.

Liu, S., Zhang, L., Long, Y., 2019. Urban Vitality Area Identification and Pattern Analysis from the Perspective of Time and Space Fusion. *SUSTAINABILITY*, 11(15).

Ma, Y., Sun, Y., Weng, F., Xu, Y., 2023. An Identification of Industrial Functional Zones Based on NLP: Evidence From Online Commercial Registration Data. *SAGE OPEN*, 13(1).

Mostafavi, S. A., Safikhani, H., Salehfard, S., 2021. Air pollution distribution in Arak city considering the effects of neighboring pollutant industries and urban traffics. *International Journal of Energy and Environmental Engineering*, 12(2), 307–333. https://doi.org/10.1007/s40095-020-00379-5.

Qiu, Q., Xie, Z., Ma, K., Tian, M., 2023. BERTCWS: unsupervised multi-granular Chinese word segmentation based on a BERT method for the geoscience domain. *Annals of GIS*, 29(3), 387–399. https://doi.org/10.1080/19475683.2023.2186487.

Sahu, N., Pipalde, D., Jain, R. C., Gupta, S. C., 2024. Geostatistical analysis of soil pollutants of tannery industrial area, Kanpur. *Indian Journal of Ecology*, 51(1), 113–117.

Tu, X., Fu, C., Huang, A., Chen, H., Ding, X., 2022. DBSCAN Spatial Clustering Analysis of Urban ”Production-Living-Ecological” Space Based on POI Data: A Case Study of Central Urban Wuhan, China. *INTERNATIONAL JOURNAL OF ENVIRONMENTAL RESEARCH AND PUBLIC HEALTH*, 19(9).

Wu, Y., Wang, H., Liu, Y., Ding, S., Wen, H., 2024. Collaborative Study on the Heavy Metal Wastewater Control and Carbon Emission of Typical Lead Zinc Smelting Enterprises in the Yellow River Basin. *Research of Environmental Sciences*, 37(1), 212–220.

Yang, L., Jin, Q., Fu, F., 2024. Research on Urban Street Network Structure Based on Spatial Syntax and POI Data. *Sustainability*, 16(5).