



OPEN

DATA DESCRIPTOR

An inventory of industrial solid waste in 337 cities of China: Applying machine learning for data completion

Qian Jia^{1,2}, Kunsen Lin^{1,3}, Jiawei Zhuang³, Dengyu Yang⁴, Wei Wei⁵, Xiong Xiao⁵, Huanzheng Du^{2,6} & Tao Wang^{1,2,7}  

Rapid industrialization of China generated a massive quantity of waste, among them industrial solid waste contributed the biggest flow to some 60 gigatonnes (Gt) in the past two decades. A complete tempo-spatial dataset of industrial waste, however, is absent in many areas in China, due to numerous waste producers and insufficient statistical coverage. To fill up the gap, we collected current available data from thousands of sources. We further developed six machine learning models to complete the dataset across all the 337 cities in China for the period 1990–2022. Bayesian optimization was employed to obtain the best estimation model for each city and to enhance its performance and resilience. In addition to the aggregate waste amount, generation of six major subcategories of industrial waste, i.e., metallurgical slags, fly ash, furnace slags, coal gangue, tailings, and desulfurization gypsum, are presented for more than half of the cities in 2022. This dataset can help researchers and policymakers recognize and address challenges brought by industrial waste.

Background & Summary

Rapid industrialization and economic development in China in the past decades have generated a colossal amount of solid waste and posed heavy risks onto the environment¹. At present waste streams from agricultural, construction, industrial, and post-use sources may aggregate to more than 10 gigatonnes per annum (Gt/a)^{2–5}. Among them, industrial solid waste (ISW) in China counts for some 4 Gt/a, which is equivalent to more than half of coal or nearly double of iron ore production in the world^{5–7}.

According to an updated solid waste code, there are 17 main categories and hundreds of subcategories of ISW in China (excluding hazardous industrial waste). The top six categories are metallurgical slags, fly ash, furnace slags, coal gangue, tailings, and desulfurization gypsum, respectively. ISW can be utilized as secondary materials, landfilled, incinerated, or temporarily stockpiled for future treatment. Although the 13th Five Year Plan of China (2016–2020) set the goal of a 73% comprehensive utilization rate, only 57.7% of ISW was reused by 2022 (2022 China's Ecological Environment Status Bulletin). The cumulative ISW stocks to be environmentally safely treated reached some 60–70 Gt⁸, led to a variety of environmental degradations from soil erosion, groundwater contamination, to habitat destruction and biodiversity loss across the country.

Despite the urgent challenge, a complete tempo-spatial dataset of ISW generation in China at city-level has not been established, preventing a comprehensive understanding and further investigation on waste management and recycling. A well-functioning environmental monitoring and data collection system can statistically track and report industrial waste from factory to factory and city by city. However, this takes time and requires expensive investment in the environmental infrastructure. Before high-resolution industrial waste information

¹College of Environmental Science and Engineering, Tongji University, Shanghai, 200092, P.R. China. ²UNEP-Tongji Institute of Environment for Sustainable Development (IESD), Shanghai, 200092, P.R. China. ³College of Environmental and Resource Sciences, Fujian Normal University, Fuzhou, 350100, Fujian, P.R. China. ⁴Shanghai Key Laboratory of Atmospheric Particle Pollution and Prevention (LAP3), Department of Environmental Science and Engineering, Fudan University, Shanghai, 200433, P.R. China. ⁵Xiaping Environmental Park, Shenzhen, 518024, P.R. China. ⁶Circular Economy Research Institute, School of Marxism, Tongji University, Shanghai, 200092, P.R. China. ⁷Institute of Carbon Neutrality, Tongji University, Shanghai, 200092, China. ✉e-mail: a.t.wang@foxmail.com

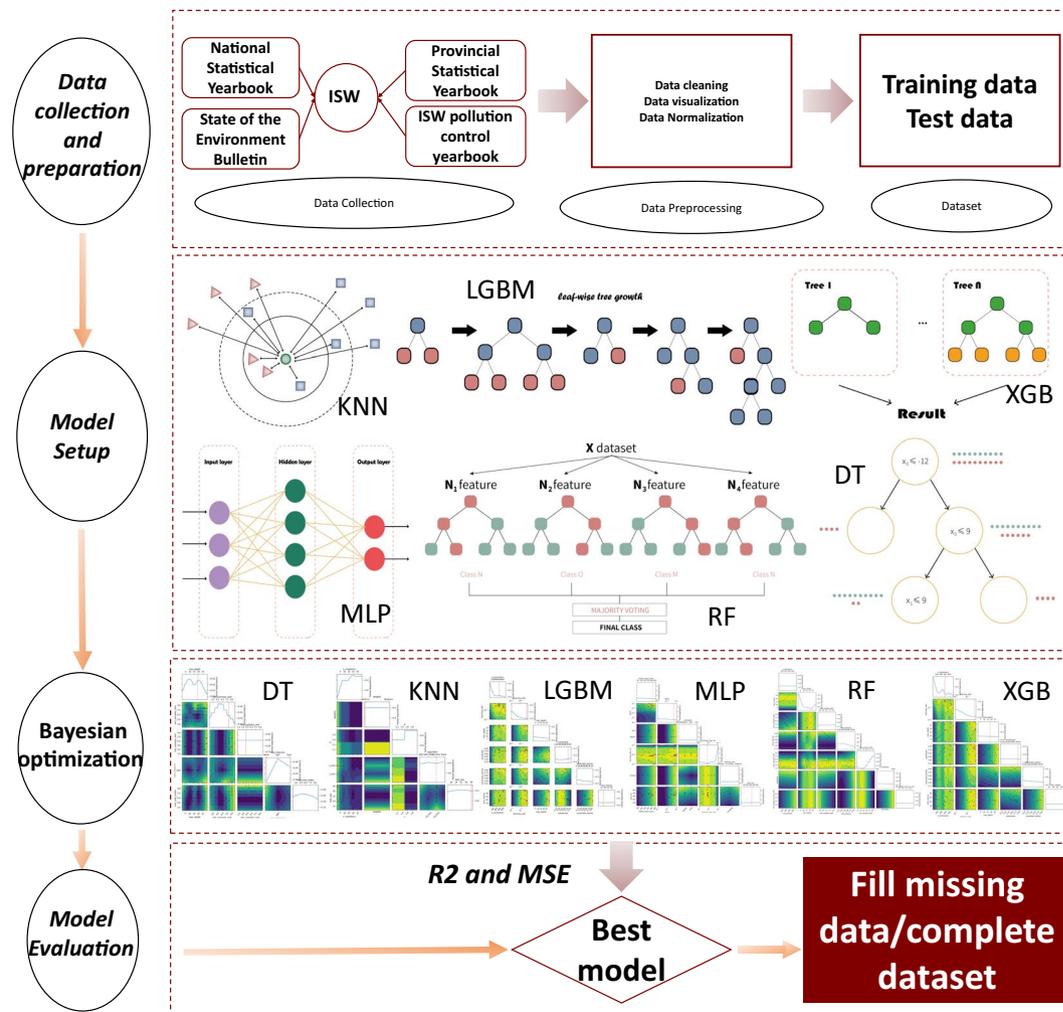


Fig. 1 Methods of ISW data completion.

are fully recorded, machine learning can provide a powerful tool for estimating the regional and sub-provincial industrial waste generation patterns in China and filling the gaps.

Machine learning has been increasingly applied in environmental management because of its cost-efficiency, predictive accuracy, and robustness. It is already widely used in making predictions, extracting features, detecting anomalies, and discovering new materials or chemicals⁹. Machine learning has also been employed in a variety of solid waste research^{10–12}. Lin *et al.*¹³ conducted a critical review on application of deep learning in solid waste management, and found that the method had been very effective in predicting waste quantity and composition. The amount of waste can be informed with one or several machine learning methods combined. Kannagara *et al.*¹⁴ employed a neural network and decision tree in modelling and estimating regional waste generation and diversion in Canada. Ma *et al.*¹⁵ also used artificial neural network to investigate the composition of municipal solid waste (MSW). Zhang *et al.*¹⁶ predicted MSW generation with five machine learning approaches.

This present research aims to establish the city-level ISW dataset in China from 1990 to 2020 by employing a non-interpolation-based machine learning approach. It covers industrial waste from all the 337 administrative divisions at the prefectural level and above in China from 1990 to 2022, including 293 prefecture-level cities, 7 prefecture-level regions, 30 autonomous prefectures, 3 prefecture-level leagues, as well as 4 municipalities (as seen in the dataset¹⁷). For simplicity, we use “city” to refer to these administrative divisions in the text below. In addition to the aggregate ISW flow, we also attempt to develop the city-level inventory of six major ISW subcategories for Year 2022, including metallurgical slags, fly ash, furnace slags, coal gangue, tailings, and desulfurization gypsum.

ISW is one of the largest and most widely spreading solid waste flows, creating environmental and ecosystem risks in many regions. Regional and meso-scale ISW generation, however, often fluctuates owing to local industrial changes and irregular operational factors. Such dynamics can hardly be captured by conventional spatial interpolation methods. Therefore, we have developed a data-driven and non-interpolation machine learning framework to enhance predictive accuracy and spatial resolution. To our best, we accomplish this complete and consistent estimation of ISW in Chinese cities. The methodology and practice offered by this present work may also help to inform waste generation patterns in the future and in other areas.

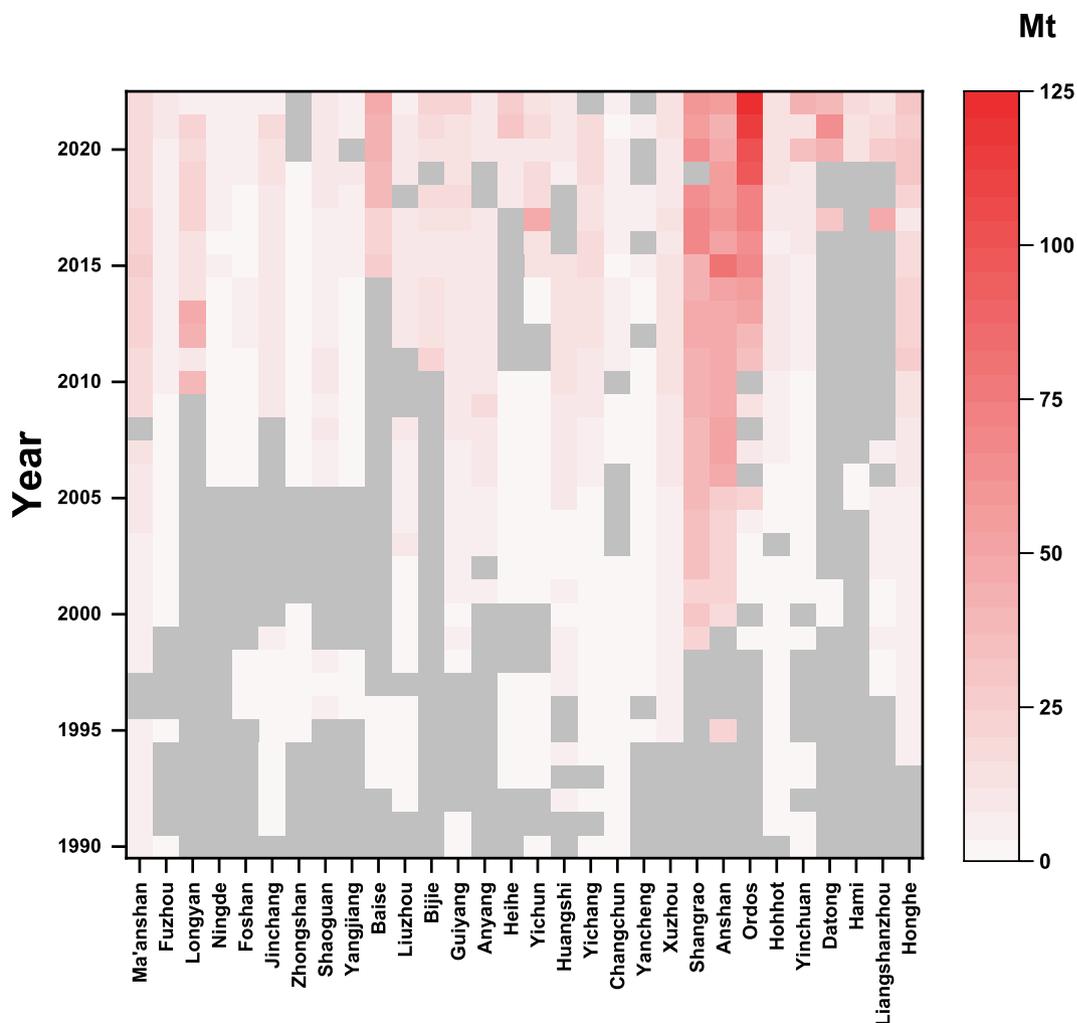


Fig. 2 Heat map of ISW in 30 top generation cities in China.

Methods

We first collect the data of ISW from available waste statistics, then set up six machine learning models and used Bayesian optimization to find the best model, and finally complete the missing data. The methodology is portrayed in Fig. 1.

Collection of ISW data. The ISW data from 337 cities across China are collected from >500 sources of national, provincial, and city statistical yearbooks and waste statistical bulletins, as depicted in the dataset¹⁷. The statistical data comprise the original ISW dataset, in which 33% are missing and to be filled. Most of these missing values occur in less developed areas or in early years, thus they are less significant in the dataset (as seen in Fig. 2).

Machine learning models. When utilizing machine learning to estimate absent data in ISW inventory, the selection of appropriate models is critical to ensure reasonable predictions. We employ six models based on their advantages and complementary effects. KNeighborsRegressor (K-Nearest Neighbour, KNN) is effective by leveraging the proximity of neighbouring data points to predict missing values. It is particularly adept at handling datasets characterized by intricate relationships and non-linear patterns¹⁸. KNN is one of the most widely-used models to estimate the generation of solid waste for its flexibility to include date type and ability to filling missing data^{19,20}. LGBMRegressor (Light Gradient Boosting Machine, LGBM) is a gradient boosting framework to make efficient and accurate computation. Notably, it is prominent in managing sizable datasets with minimal risk of overfitting²¹, as well as in identifying complex connections and capturing significance of features²². RandomForestRegressor (Random Forest, RF) employs an ensemble learning approach, constructs multiple decision trees and amalgamates their predictions. It exhibits resilience, and can adeptly manage non-linear relationships while mitigating the risk of overfitting²³. RF can predict the generation of waste even with limited data²⁴. MLPRegressor (Multilayer Perceptron, MLP) represents a form of artificial neural network esteemed for its proficiency in discerning complex data patterns, making it well-suited for datasets characterized by non-linear relationships and large data volumes²⁵. Extreme Gradient Boosting (XGB) Regressor presents another gradient boosting approach enabling great efficiency, accuracy, and resilience against outliers in regression tasks²⁶. Decision Trees (DT) offer a straightforward yet potent modelling technique, dividing data based on feature

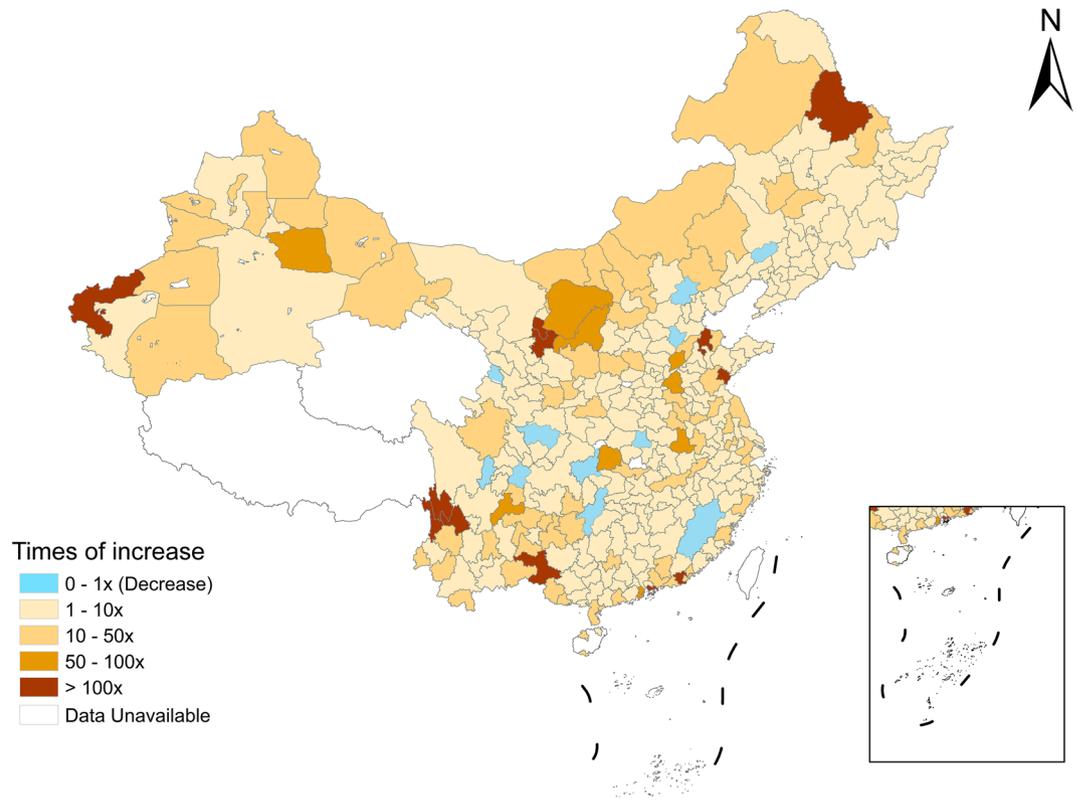


Fig. 3 Increase of ISW in 1990–2022 in cities of China.

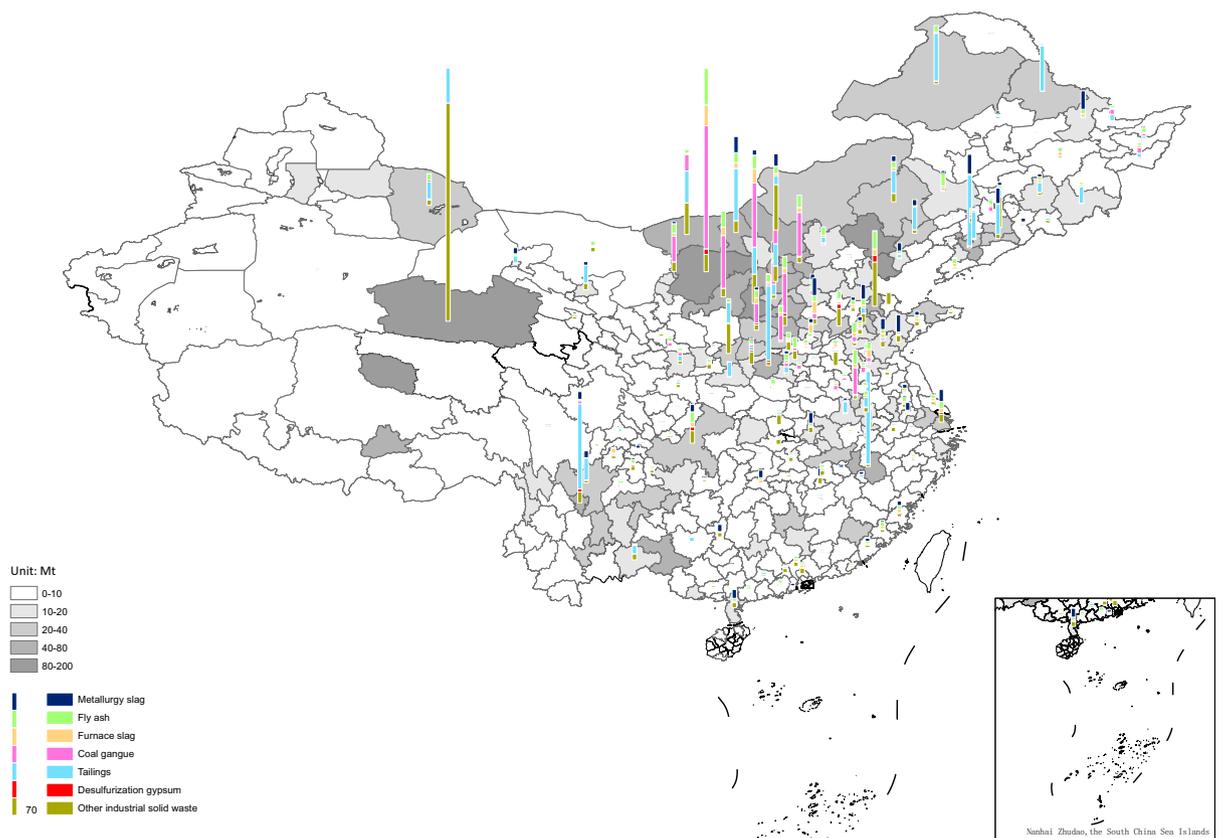


Fig. 4 Generation of ISW in cities of China in 2022, the grey area indicates ISW generation in total, and the colored column represents six subcategories of ISW.

Provinces (No. of cities)	Metrics	DT	KNN	LGBM	MLP	RF	XGB
Anhui (16)	Adjusted R2_test	0.782	0.718	0.964	/	0.661	0.955
	MSE_test	15731	20275	2565	/	24420	3224
Fujian (9)	Adjusted R2_test	0.137	0.847	0.354	/	0.617	0.947
	MSE_test	83706	14864	63684	/	37110	5105
Gansu (14)	Adjusted R2_test	0.804	0.898	0.466	-0.194	0.965	0.870
	MSE_test	20127	10418	54718	122294	3599	13330
Guangdong (21)	Adjusted R2_test	0.467	0.824	0.484	0.467	0.924	0.960
	MSE_test	28108	9297	27211	28107.700	3987	2087
Guangxi (14)	Adjusted R2_test	0.952	0.981	0.173	0.198	0.985	0.984
	MSE_test	22671	9129	392541	380296	6923	7647
Guizhou (9)	Adjusted R2_test	-0.100	-0.077	-0.093	-0.251	-0.072	-0.059
	MSE_test	42835	41936	42562	48716	41757	41249
Hebei (11)	Adjusted R2_test	0.942	0.969	0.746	0.055	0.987	0.962
	MSE_test	675070	359826	2942654	10967061	152356	443062
Henan (14)	Adjusted R2_test	0.836	0.919	0.251	0.429	0.942	0.918
	MSE_test	60971	30046	278623	212262	216034	30670
Heilongjiang (13)	Adjusted R2_test	0.172	0.929	0.402	-0.560	0.940	0.955
	MSE_test	75787	6518	54714	142681	5447	4112
Hubei (13)	Adjusted R2_test	0.649	0.866	0.659	0.035	0.858	0.955
	MSE_test	71609	27251	69444	196736	28991	9228
Hunan (14)	Adjusted R2_test	0.659	0.759	0.740	0.012	0.898	0.933
	MSE_test	28260	20007	21590	81976	8439	5522
Jilin (9)	Adjusted R2_test	/	0.882	/	0.154	0.913	0.907
	MSE_test	/	13995	/	100431	10355	10990
Jiangsu (13)	Adjusted R2_test	/	0.960	/	-0.006	0.987	0.991
	MSE_test	/	15889	/	400932	5208	3394
Jiangxi (11)	Adjusted R2_test	0.975	0.993	/	0.811	0.994	0.992
	MSE_test	48807	13879	/	375865.500	11364	15768
Liaoning (14)	Adjusted R2_test	/	0.633	-0.143	0.000	0.892	0.841
	MSE_test	/	558383	1736208	1519475	163993	241168
Inner Mongolia (12)	Adjusted R2_test	0.502	0.833	0.067	0.414	0.801	0.818
	MSE_test	1376167	462105	2576858	1619270	549723	503258
Ningxia (5)	Adjusted R2_test	0.814	0.963	0.499	-0.695	0.958	0.945
	MSE_test	16173	3233	43631	147584	3630	4773
Shandong (16)	Adjusted R2_test	/	0.939	/	0.063	0.974	0.975
	MSE_test	/	36989	/	564314	15594	15235
Shanxi (11)	Adjusted R2_test	0.511	0.674	0.335	-0.136	0.886	0.817
	MSE_test	308339	205553	418711	715351	71612	115227
Shaanxi (10)	Adjusted R2_test	0.936	0.527	0.807	0.224	0.985	0.954
	MSE_test	38179	282946	115464	464846	9242	27301
Sichuang (21)	Adjusted R2_test	0.948	0.951	0.864	-0.013	0.966	0.915
	MSE_test	21117	19727	55043	409693	13615	34183
Xinjiang (14)	Adjusted R2_test	0.829	0.955	0.586	0.298	0.858	0.944
	MSE_test	65959	17238	159222	270054	54475	21478
Yunnan (16)	Adjusted R2_test	0.092	0.792	0.570	0.368	0.923	0.944
	MSE_test	418064	95843	198035	291096	35414	25995
Zhejiang (11)	Adjusted R2_test	0.895	0.973	0.831	0.555	0.982	0.964
	MSE_test	6992	1805	11254	29613	1164	2412

Table 1. Prediction performance of test data.

values. It is known for interpretability, versatility in handling both numerical and categorical data, and efficacy in capturing non-linear relationships²⁷. A mix of different models to estimate solid waste generation or its features have been applied in a number of studies^{28,29}.

Bayesian optimization. To enhance the predictive model's performance and resilience, Bayesian optimization has been integrated alongside the selected models. It helps to tune model hyperparameters, conduct efficient exploration, and withstand noise interference. This technique aids in refining the hyperparameters of the machine

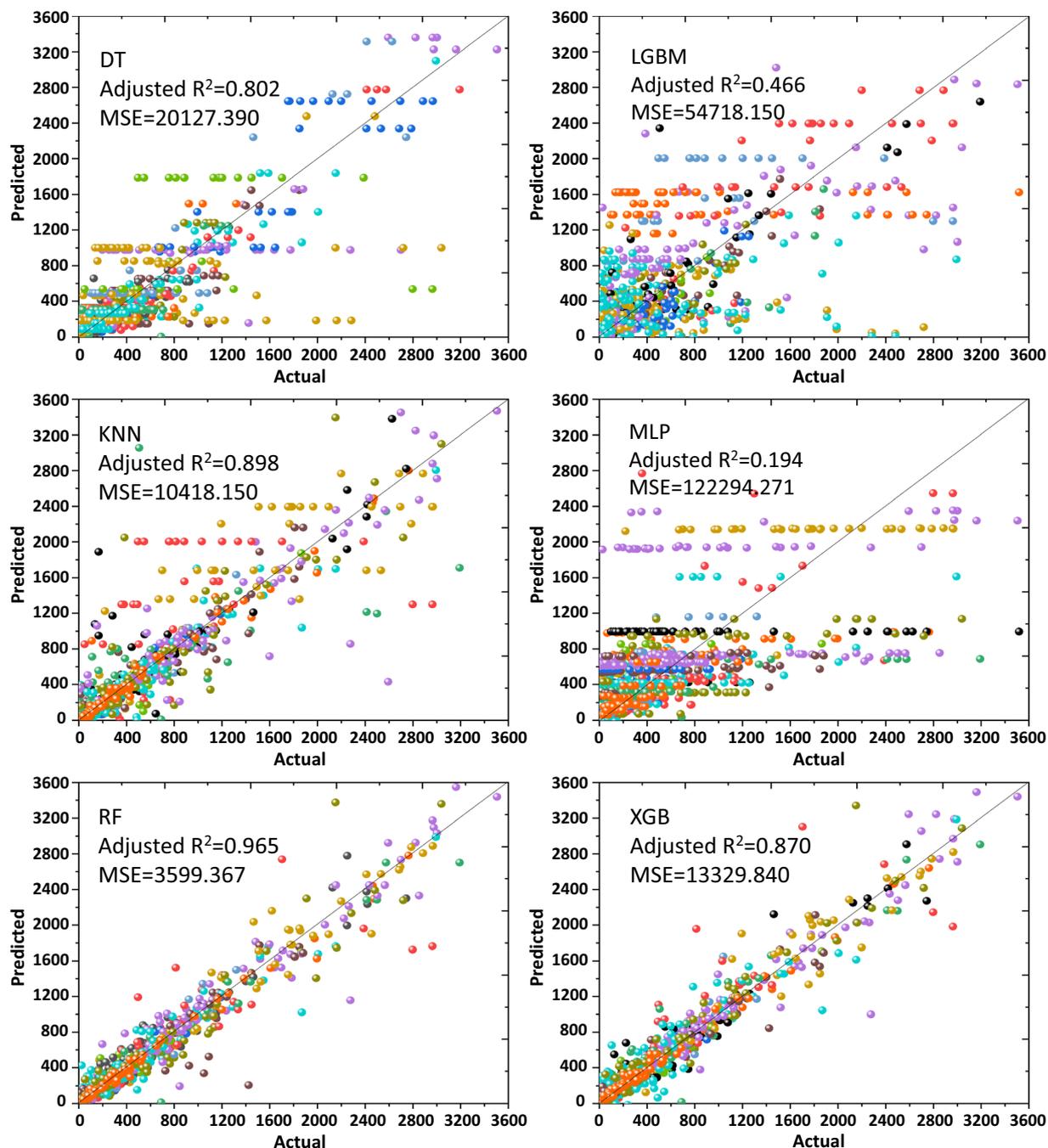


Fig. 5 Predicted values and actual values of the testing set of machine learning models, dots in the same colour represent cities from the same province.

learning models, and holds significant influence over the model's performance, including learning rate, number of estimators, and tree depth. Bayesian optimization systematically explores the hyperparameter space to pinpoint the optimal configuration, consequently refining model accuracy and generalization. By employing probabilistic models, it arranges the search for optimal hyperparameters, maintaining a delicate balance between exploring new regions in the hyperparameter space and exploiting promising areas, thereby ensuring efficient and effective optimization. This strategy of exploration and exploitation is pivotal, especially when grappling with intricate models and datasets containing incomplete information³⁰. Moreover, Bayesian optimization exhibits robustness in the face of noise when evaluating objective functions. In scenarios where machine learning models are trained on incomplete data, the presence of missing values can introduce disruptive noise in the optimization process. By encapsulating the uncertainty linked with the objective function, Bayesian optimization mitigates the impact of noise, yielding more dependable hyperparameter estimates³¹.

Moreover, Bayesian optimization can automate the process of hyperparameter tuning, reducing the burden on data scientists and practitioners³². By iteratively selecting hyperparameters based on past evaluations,

Bayesian optimization efficiently converges to the optimal solution. This automation accelerates the model development pipeline and enables faster experimentation with different machine learning algorithms and configurations.

Data Records

“Industrial solid waste dataset in China 1990–2022” can be accessed under Figshare¹⁷. It includes the following data sheets:

- List of prefectural cities, showing 333 prefectural level administrative divisions across 27 provincial level divisions in mainland China. It includes an inventory of 293 prefecture-level cities, 7 prefecture-level regions, 30 autonomous prefectures, 3 prefecture-level leagues.
- ISW at city level (original), presenting the original data of ISW generation in 333 prefectural divisions and 4 municipalities in China in the period 1990–2022. The proportion of missing data is 33%.
- ISW at city level (computed), presenting the data that are computationally developed in 333 prefectural divisions and 4 municipalities in China in the period 1990–2022.
- ISW at city level (complete), presenting the complete data of ISW generation in 333 prefectural divisions and 4 municipalities in China in the period 1990–2022. The missing value in the original dataset has been completed with the best estimate of the machine learning models.
- Six major ISW at city level, presenting the original data of the generation of six subcategories of ISW in 173 prefectural divisions and 3 municipalities in China in 2022, including metallurgical slags, fly ash, furnace slags, coal gangue, tailings, and desulfurization gypsum.
- Data sources, listing the sources of the original data in the sheets of “ISW at city level (original)” and “Six major ISW at city level”.

Based upon the complete ISW dataset, Fig. 3 shows the increase rate of ISW in 2022 of all Chinese cities compared to that in 1990. Cities in resource rich areas or economically advanced areas showed the biggest increase in ISW. This suggests that ISW generation was largely driven by industrial development coupled with rapid urbanization. From the 1990s, industrialization surged as the government encouraged investment in manufacturing and infrastructure, leading to the emergence of bustling urban centers and the expansion of industrial zones across the country. Concurrently, urbanization accelerated as rural residents migrated to cities in search of employment opportunities. Since the beginning of reform and opening up in the late 1970s, the industrial value-added in China had increased nearly 200 times until 2022³³, and urbanization rate had climbed from 17.92% to 65.2%³⁴. The ISW generation, nonetheless, grew by a factor of 8.4 in the period of 1980 to 2022.

Specifically, among all ISW flows, metallurgical slags, fly ash, furnace slags, coal gangue, tailings, and desulfurization gypsum are the major subcategories in China, coded as SW01 to SW06, respectively, under the recently updated classification of the Ministry of Environment and Ecology of China³⁵. They are produced alongside with industrial activities including mining, metallurgy, thermal power industry. Due to data limitation, time series dataset for subdivision of ISW cannot be provided by far. We compiled generation of six ISW subcategories in major Chinese cities in 2022, as illustrated in Fig. 4 and presented in the dataset¹⁷.

Technical Validation

Model evaluation. To assess the effectiveness of these models, two metrics were utilized: the correlation coefficient (R^2) and mean square error (MSE). These metrics offer insights into the precision and predictive capability of the models. R^2 quantifies the strength and direction of the linear association between predicted and actual values, ranging from 0 to 1, where 1 signifies a perfect correlation. MSE gauges the average squared disparity between predicted and actual values, offering a comprehensive assessment of the model’s precision, with lower values indicating superior performance. Therefore, R^2 and MSE are calculated with the following equations, respectively.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

where y_i , \hat{y}_i , and \bar{y} represented the actual, predicted and average value of ISW, respectively. The values of R^2 and MSE for each model are shown in Table 1.

To further validate the data quality, we compared predicted data with actual data of all the cities in year 2022. Figure 5 shows the scatter plot of predicted values and actual values of testing set, which indicates the performance of different models in prediction. The bolded table cell in Table 1 represents the model applied to certain provinces and cities, with the best performance (R^2 closes to 1 and MSE the smallest among all models). Finally, the missing ISW values are completed by the best prediction models for each city.

Limitations and perspectives. This study attempts to complete an ISW dataset for across 337 cities in China as from 1990 to 2022, yet certain limitations remain. The quality and consistency of the data input, retrieved primarily from statistical yearbooks and waste bulletins and reports, may affect the predictive accuracy due to regional and temporal inconsistencies. Especially in small cities, even opening or closure of a coal mine or

steel plant may substantially change their ISW generation. This business change has randomness, thus is uneasily predicted by machine learning models. Additionally, the feature selection process does not fully account for complex socio-economic factors such as policy shift, business movement, and technological advancement, all could significantly change ISW patterns. The models also demonstrate limited adaptability in regions with sparse or highly variable data, as evidenced by lower R^2 scores or even negative values in some cities and provinces. Furthermore, the study predominantly focuses on static predictions based on historical data, offering limited insights into the dynamic evolution of waste toward future, including increasingly rigorous environmental regulations and adoption of innovative green technologies.

Despite these constraints, the findings lay a solid foundation for enhancing waste management research. Future studies could integrate real-time monitoring data obtained from IoT and remote sensing to improve data coverage and accuracy. Incorporating advanced deep learning frameworks, such as graph neural networks or transformer architectures, could improve the models' capability to handle spatial and temporal complexities. Scenario-based dynamic modelling, combined with interdisciplinary approaches, could provide a deeper understanding of ISW generation trends and their environmental implications, enabling policymakers and industrial stakeholders to develop more effective waste management strategies and promote sustainable and low-carbon development pathways.

This dataset of ISW generation across prefectural divisions and municipalities in China holds potential for diverse applications. It can constitute a foundation to estimate future waste streams, identify regional disparities, and evaluate waste reduction programs. Furthermore, the dataset can support the development of predictive models for solid waste under various economic and regulatory scenarios. Policymakers and environmental agencies may also leverage these data to design targeted interventions, optimize resource recycling, and promote circular economy initiatives. Future research may integrate this dataset with socioeconomic, environmental, and industrial structure indicators and yield more informative insights into waste management and recycling in specific regions.

Code availability

Data were processed with Python. The code has been uploaded to https://github.com/MLforSW/ISWM_Imputation.

Received: 14 March 2025; Accepted: 10 July 2025;

Published online: 16 July 2025

References

1. Yang, Z., Chen, H., Du, L., Lu, W. & Qi, K. Exploring the industrial solid wastes management system: Empirical analysis of forecasting and safeguard mechanisms. *J. Environ. Manage.* **279**, 111627, <https://doi.org/10.1016/j.jenvman.2020.111627> (2021).
2. Chen, J. Current status and suggestions for the recycling of construction waste in China. *Constr. Sci. Technol.* **1**, 8–12 (2014).
3. Food and Agriculture Organization of the United Nations (FAO). *Food loss and waste database*. <https://www.fao.org/platform-food-loss-waste/flw-data/en> (2025).
4. Li, C. *et al.* A systematic review of food loss and waste in China: Quantity, impacts and mediators. *J. Environ. Manage.* **303**, 114092, <https://doi.org/10.1016/j.jenvman.2021.114092> (2022).
5. National Bureau of Statistics of China (NBSC). *China Statistical Yearbook 2023*. China Statistics Press (2023).
6. International Energy Agency (IEA). *World Energy Statistics*. <https://www.iea.org/data-and-statistics/data-product/world-energy-statistics> (2024).
7. World Steel Association (WSA). *2023 World Steel in Figures*. WSA (2023).
8. Zhang, X. *et al.* Analysis of driving factors on China's industrial solid waste generation: Insights from critical supply chains. *Sci. Total Environ.* **775**, 145185, <https://doi.org/10.1016/j.scitotenv.2021.145185> (2021).
9. Zhong, S. *et al.* Machine learning: New ideas and tools in environmental science and engineering. *Environ. Sci. Technol.* **55**, 12741–12754 (2021).
10. Alidoust, P. *et al.* Prediction of the shear modulus of municipal solid waste (MSW): An application of machine learning techniques. *J. Clean. Prod.* **303**, 127053, <https://doi.org/10.1016/j.jclepro.2021.127053> (2021).
11. Andeobu, L., Wibowo, S. & Grandhi, S. Artificial intelligence applications for sustainable solid waste management practices in Australia: A systematic review. *Sci. Total Environ.* **834**, 155389 (2022).
12. You, H. *et al.* Comparison of ANN (MLP), ANFIS, SVM, and RF models for the online classification of heating value of burning municipal solid waste in circulating fluidized bed incinerators. *Waste Manage.* **68**, 186–197, <https://doi.org/10.1016/j.wasman.2017.03.044> (2017).
13. Lin, K. *et al.* Toward smarter management and recovery of municipal solid waste: A critical review on deep learning approaches. *J. Clean. Prod.* **346**, 130943, <https://doi.org/10.1016/j.jclepro.2022.130943> (2022).
14. Kannangara, M., Dua, R., Ahmadi, L. & Bensebaa, F. Modeling and prediction of regional municipal solid waste generation and diversion in Canada using machine learning approaches. *Waste Manage.* **74**, 3–15, <https://doi.org/10.1016/j.wasman.2017.11.057> (2018).
15. Ma, S., Zhou, C., Chi, C., Liu, Y. & Yang, G. Estimating physical composition of municipal solid waste in China by applying artificial neural network method. *Environ. Sci. Technol.* **54**, 9609–9617, <https://doi.org/10.1021/acs.est.0c01802> (2020).
16. Zhang, C., Dong, H., Geng, Y., Liang, H. & Liu, X. Machine learning based prediction for China's municipal solid waste under the shared socioeconomic pathways. *J. Environ. Manage.* **312**, 114918, <https://doi.org/10.1016/j.jenvman.2022.114918> (2022).
17. Jia, Q. *et al.* An inventory of industrial solid waste in 337 cities of China: Applying machine learning for data completion. *figshare* <https://doi.org/10.6084/m9.figshare.28504682> (2025).
18. Daniels, A. & Koutsougeras, C. Predicting water quality parameters in Lake Pontchartrain using machine learning: A comparison on K-nearest neighbors, decision trees, and neural networks to predict water quality. *Proc. 5th Int. Conf. Inf. Syst. Data Min.* 28–33, <https://doi.org/10.1145/3471287.3471308> (2021).
19. Pal, A., Ahmed, K. S., Hossain, F. M. Z. & Alam, M. S. Machine learning models for predicting compressive strength of fiber-reinforced concrete containing waste rubber and recycled aggregate. *J. Clean. Prod.* **423**, 138673, <https://doi.org/10.1016/j.jclepro.2023.138673> (2023).
20. Working Committee of Construction and Demolition Waste Management and Recycling, China Association of Urban Environmental Sanitation. Personal communication (October 2024).
21. Chen, H. *et al.* Shield attitude prediction based on Bayesian-LGBM machine learning. *Inf. Sci.* **632**, 105–129, <https://doi.org/10.1016/j.ins.2023.03.004> (2023).

22. Dong, Q. *et al.* A city-level dataset of heavy metal emissions into the atmosphere across China from 2015–2020. *Sci. Data* **11**, 258, <https://doi.org/10.1038/s41597-024-03089-3> (2024).
23. Kamal, N. & Raheja, S. Prediction of software vulnerabilities using random forest regressor. In *Computational Intelligence* (eds Shukla, A., Murthy, B. K., Hasteer, N. & Van Belle, J.-P.) 411–424, https://doi.org/10.1007/978-981-19-7346-8_35 (2023).
24. Cha, G.-W. *et al.* Development of a prediction model for demolition waste generation using a random forest algorithm based on small datasets. *Int. J. Environ. Res. Public Health* **17**, 6997 <https://www.mdpi.com/1660-4601/17/19/6997> (2020).
25. Massaoudi, M. *et al.* A novel stacked generalization ensemble-based hybrid LGBM-XGB-MLP model for short-term load forecasting. *Energy* **214**, 118874, <https://doi.org/10.1016/j.energy.2020.118874> (2021).
26. Loś, H., *et al.* Evaluation of XGBoost and LGBM performance in tree species classification with Sentinel-2 data. In *2021 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)* 5803–5806, <https://doi.org/10.1109/IGARSS47720.2021.9553031> (2021).
27. Charbuty, B. & Abdulazeez, A. Classification based on decision tree algorithm for machine learning. *J. Appl. Sci. Technol. Trends* **2**, 01, <https://doi.org/10.38094/jastt20165> (2021).
28. Nguyen, X. C. *et al.* Development of machine learning-based models to forecast solid waste generation in residential areas: a case study from Vietnam. *Resour. Conserv. Recycl.* **167**, 105381, <https://doi.org/10.1016/j.resconrec.2020.105381> (2021).
29. Qi, Y.-P., He, P.-J., Lan, D.-Y., Lü, F. & Zhang, H. Novel method for predicting concentrations of incineration flue gas based on waste composition and machine learning. *J. Environ. Manage.* **373**, 123588, <https://doi.org/10.1016/j.jenvman.2024.123588> (2025).
30. Du, L., Gao, R., Suganthan, P. N. & Wang, D. Z. W. Bayesian optimization based dynamic ensemble for time series forecasting. *Inf. Sci.* **591**, 155–175, <https://doi.org/10.1016/j.ins.2022.01.010> (2022).
31. Nguyen, V.-H. *et al.* Applying Bayesian optimization for machine learning models in predicting the surface roughness in single-point diamond turning polycarbonate. *Math. Probl. Eng.* **2021**, 6815802, <https://doi.org/10.1155/2021/6815802> (2021).
32. Yang, K., Liu, L. & Wen, Y. The impact of Bayesian optimization on feature selection. *Sci. Rep.* **14**, 3948, <https://doi.org/10.1038/s41598-024-54515-w> (2024).
33. Yuan, J. *et al.* Ecology of industrial pollution in China. *Ecosyst. Health Sustain.* **6**, 1779010 (2020).
34. Wang, Y. & Wang, L. New-type urbanization and rural revitalization: A study on the coupled development of the Yangtze River Economic Belt, China. *PLoS ONE* **20**, e0314724 (2025).
35. Ministry of Ecology and Environment (MEE). *Solid Waste Classification and Code Directory 2024*. MEE (2024).

Acknowledgements

This research was supported by the National Key Research and Development Program of China (2022YFC3803401), the National Natural Science Foundation of China (No. 71974144), the Research and Education Project for Young and Mid-career Faculty in Fujian Province, China (Jz24011), the research fund of Xiaping Environmental Park of Shenzhen City, and the open fund of the Institute of Carbon Neutrality, Tongji University. We gratefully thank the anonymous reviewers and the editor. Their valuable comments and suggestions helped us to improve the quality of the work.

Author contributions

Q.J. developed the dataset and drafted the manuscript. K.L. conducted modelling, made technical validation, and revised the manuscript. J.Z. and D.Y. contributed to the modelling and technical validation. W.W. and X.X. collected and compiled the data. H.D. led the project and supervised the research. T.W. supervised the research, compiled the data, and revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to T.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025