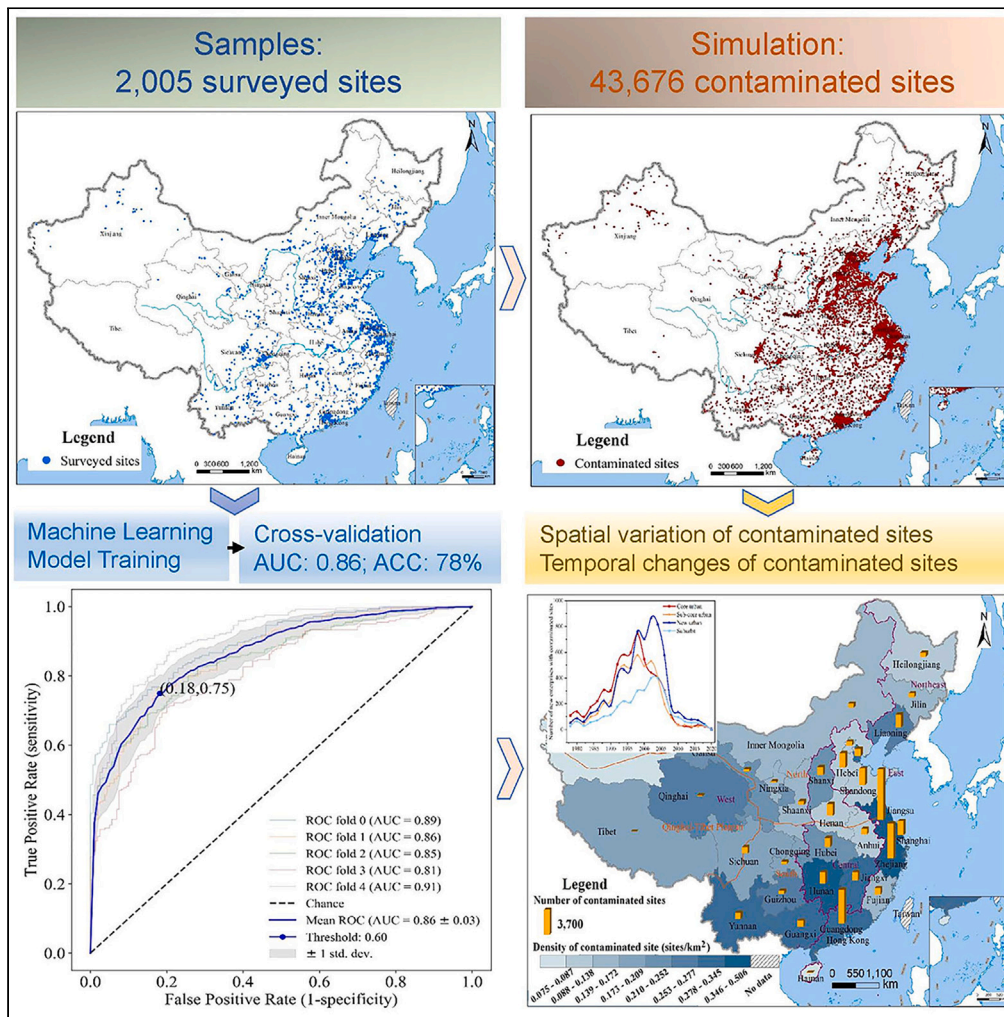## Article

The rapid increase of urban contaminated sites along China's urbanization during the last 30 years

Kai Li, Ranhao Sun, Guanghui Guo

rhsun@rcees.ac.cn

Highlights

Machine-learning models can identify contaminated sites efficiently and quickly

Urban contaminated sites in China have significant spatiotemporal heterogeneity

Our findings provide insights into accurate and efficient environmental policies

Article

The rapid increase of urban contaminated sites along China's urbanization during the last 30 years

Kai Li,[1,2] Ranhao Sun,[1,2,4,*] and Guanghui Guo[2,3]

SUMMARY

Contaminated sites pose serious threats to the soil environment and human health. However, the location and temporal changes of urban contaminated sites across China remain unknown due to data scarcity. Here, we developed a machine-learning model to identify the contaminated sites using public data. Results show that the trained model with 2,005 surveyed site samples and six variables can achieve a model performance evaluation value of 0.86. 43,676 contaminated sites were identified from 83,498 polluting enterprise plots in China. However, these contaminated sites have significant spatiotemporal heterogeneity, mainly located in economically developed provinces, urban agglomerations, and core urban areas. Moreover, the contaminated sites increased by 325% along with urban expansion from 1990 to 2018. The abandoned contaminated sites increased rapidly, but the contaminated sites in production decreased continuously. This methodological framework and our findings contribute to the precise management of contaminated sites and provide insights into urban sustainable development.

INTRODUCTION

Over the past three decades, China has been one of the world's fastest urbanizing and industrializing countries.[1,2] China overtook the US as the world's largest urban area in 2015, and the world's largest manufacturing output in 2011.[3] As the urban area has expanded and restructured, more and more polluting enterprises with contaminated sites have been closed, suspended operations, merged, or relocated.[4,5] On the one hand, contaminated sites threaten the soil ecology and reduce the quality of groundwater. According to the National Soil Pollution Survey Bulletin published in 2014, 36% and 35% of soil survey points in China's heavily polluted enterprise sites and industrial waste sites, respectively, exceeded the threshold.[6] On the other hand, they pose a serious risk to human health,[7] reduce the value of land, and hamper urban regeneration.[8] For example, in 2015, a chemically contaminated site around the Changzhou Foreign Language School caused hundreds of students to suffer physical abnormalities. Therefore, the environmental management of contaminated sites caused by urbanization has become a pressing issue facing China.[9]

Although the positive effects of urbanization on economic development and technological innovation have been demonstrated,[10] much work has also documented the negative ecological and environmental impacts of urban expansion and its spatial variation.[11,12] Firstly, urbanization dramatically changes the type of land use and land cover from agricultural to buildup,[13,14] directly leading to the loss of cultivated land and threatening biodiversity through habitat destruction and fragmentation.[12,15] Then newly added large impermeable surface creates an urban heat island and changes climate by affecting local temperature, humidity, and air convection.[16–19] In addition, urbanization increases the concentration of air pollutants to form haze,[20] reduces water quality,[21] and creates contaminated sites by polluting soil and groundwater.[22] The investigation, remediation techniques, and environmental management of contaminated site lag behind water and air in China due to their concealment, irreversibility, and accumulation.[23] Therefore, much less is known about the location of contaminated sites and their characterization of spatiotemporal patterns during urbanization in China.

At the site scale, researchers have made important advances in the contaminated site investigation methods,[24] risk assessment systems, and remediation technologies.[25–27] As the identification of contaminated sites requires complex and expensive investigation and assessment, contaminated site management is often one of the most technically challenging and costly sites.[28] There is generally an initial assessment of potentially contaminated sites using a prioritization approach, and the results determine whether further investigation is required to confirm the presence of site contamination.[29] Drawing on the experience of the US Superfund HRS (Hazard Ranking System) relative risk assessment,[30] environmental managers in China have established a risk screening system to rank the environmental risks of about 110,000 enterprise plots through field surveys to obtain detailed information.[31] With limited resources, this approach is useful for improving management efficiency

[1]State Key Laboratory of Urban and Regional Ecology, Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, Beijing 100085, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
[3]Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China
[4]Lead contact
*Correspondence: rhsun@rcees.ac.cn
https://doi.org/10.1016/j.isci.2023.108124

Table 1. T test and chi-squared test of potential variables with dependent variables

Potential variables	T test	Potential variables	Chi-squared test
Duration	20.29[a]	Starting time	401.96[a]
Violations	12.20[a]	Scale	6.41
Pollutant mobility	0.69	Impervious surfaces	1.03
Pollutant volatility	1.95	Industry class	220.48[a]
Soil texture	0.09	Persistent organic pollutants	0.51
Precipitation	7.40[a]	Soil erosion	5.85
Temperature	3.36[a]		
Wind speed	1.50		

[a]mean $p < 0.05$.

and controlling the risk of contaminated sites. However, the high cost of using these methods makes it difficult to identify all the contaminated sites on millions of enterprise plots and to study their spatial characteristics in China.

To fill this gap, the machine-learning model was used here for the identification of contaminated sites in China, which has a powerful computational capability to quickly extract unknown information from sample data.[32] Currently, machine-learning algorithms are widely used in the field of environmental pollution detection with high accuracy. For example, a machine-learning model was used to identify and classify potential polluters in the Yangtze River Delta region of China with an accuracy of 87%.[33] To estimate the spatial distribution of heavy metals in soil more accurately, a backpropagation (BP) neural network was used, which reduced the error by 42% compared to Kriging interpolation.[34] In groundwater, the random forest (RF) model predicted groundwater fluoride contamination across India with an accuracy of 0.78% at a resolution of 1 km.[35] In addition, the increase in the number of site surveys in recent years has provided a large number of training samples for the machine-learning algorithm to identify contaminated sites.

Under such conditions, this study utilized machine-learning models to identify contaminated sites from polluting enterprise plots and investigated the spatiotemporal patterns of contaminated sites in the process of urbanization in China, based on the data of 83,498 polluting enterprises and 2,005 surveyed sites. We find that the binary logistic regression (BLR) machine-learning model can accurately identify 43,676 contaminated sites. However, these contaminated sites were widely and unevenly distributed across China. Massive urban expansion has led to a rapid increase in contaminated sites from 1990 to 2018. This new knowledge of contaminated sites in China is remarkable for managers to formulate accurate and efficient environmental policies, promote urban development, and improve our understanding.

RESULTS

Model results and accuracy of contaminated sites

Input variable screening

The t test and chi-squared test were used to screen six variables from 14 potential variables, including duration, starting time, industry class, violations, precipitation, and temperature (Table 1). These variables are significantly correlated with the dependent variables ($p < 0.05$), and the variance inflation factors are less than 10 (Table S1), indicating that there is no multicollinearity between them.[36]

To validate the accuracy of variable selection, BF and BLR models were constructed separately using the 14 variables to obtain variable importance, coefficients, and accuracy. As shown in Table S2, the top six variables ranked by importance score in the RF model are consistent with the results of the t test and chi-square analysis. Except for soil texture, the top six variables ranked by the absolute value of coefficients in the BLR model are also consistent with the results of the t test and chi-square analysis. Therefore, the selection results of input factors are accurate and reliable.

Optimal machine-learning model

The receiver operating characteristic (ROC) curve and accuracy value (ACC) were used to evaluate the performance of machine-learning model in this study. The area under the ROC curve, known as the AUC value, was utilized to assess the overall performance of the model. 2,005 samples containing 6 variables were input into the BLR, RF, BP, and support vector machine (SVM) models to train the best-performance model evaluated with AUC value through 5-fold cross-validation. As shown in Table 2, the AUC values of the BLR, BP, and SVM models are the highest, reaching 0.86. The highest ACC values are observed for the BLR and SVM models, both achieving 0.78. The BP model exhibits the highest sensitivity value of 0.78, while the BLR model demonstrates the highest specificity value of 0.82. Therefore, the BLR model was the best performer in three of the four performance metrics, making it the majority performer. Compared to the RF, BP, and SVM models, the BLR model is less complex, more computationally efficient, and provides the coefficients of the variables, making it the chosen model for identifying contaminated sites. Furthermore, compared to the accuracy of the model before variable selection, the selected variables improved the accuracy of the RF model by 0.03, while the BLR model remained unchanged. This further confirms the accuracy and scientific validity of variable selection.

Table 2. Comparison table of machine-learning model evaluation metrics

Machine-learning model	AUC	ACC	Sensitivity value	Specificity value	Threshold value
BLR	0.86	0.78	0.75	0.82	0.60
RF	0.83	0.77	0.68	0.81	0.71
BP	0.86	0.77	0.78	0.78	0.56
SVM	0.86	0.78	0.76	0.81	0.59

The trained BLR model achieved an average AUC value of 0.86 across the five validation sets (Figure 1A) and an accuracy rate of 78% using a threshold value of 0.60, which was determined by the highest value of the Youden index (Figure 1B). The difference between the minimum AUC of 0.81 and the maximum AUC of 0.91 was small and remained at a high level over five verifications, indicating stable model performance (Figure 1A). The best trade-off between sensitivity and specificity was achieved with a threshold value of 0.60. The sensitivity value of 0.75 and the specificity value of 0.82 indicate that the trained model can correctly identify 75% of the contaminated sites and 82% of the uncontaminated sites (Figure 1B).

Figure 1C shows the ranking of the influence coefficient of the input variables in the multivariate logistic regression model. Duration, starting time, industry class, violations, and precipitation are positively correlated with the probability of site contamination, while temperature is negatively correlated. Duration is the most important variable in the multivariate logistic regression model. In addition, industry class and violations play a moderate role. Starting time, precipitation, and temperature play a low role.

Model simulation

The trained model was applied to simulate 83,498 enterprise plots in the urban area. The content of soil and groundwater pollutants in 43,676 enterprise plots exceeded the national standard value, covering a total of 52.31% of the country. Since the trained model can identify 75% of the contaminated sites, the 43,676 enterprise plots with a high pollution probability theoretically contain 75% of the contaminated sites that can be used to study contaminated sites in China's urban area.

As can be seen in Table 3, there are large differences in the number and proportion of contaminated sites by industry division. The three industries with the highest number of contaminated sites are the chemical (CM), electroplating (ET), and non-ferrous metal smelting (NFMS), while the three industries with the highest proportion of contaminated sites are chemical APIs, papermaking, and non-ferrous metal mining and dressing (NFMD).

Spatial variation of contaminated sites

Contaminated sites are very widespread and unevenly distributed across China, with significant spatial variation. The number and density of contaminated sites by province varies from 29 to 7,405 sites and from 0.08 to 0.51 sites per square kilometer (Figure 2). China's contaminated
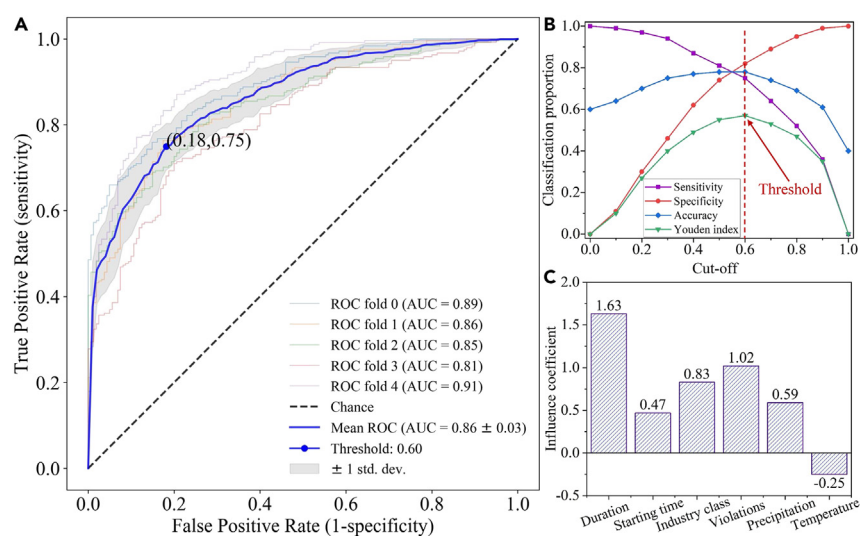

Figure 1. BLR modeling results
(A) ROC curve.
(B) Sensitivity, specificity, accuracy, and Youden index plotted against cutoff.
(C) Influence coefficient of input variables in BLR.

Table 3. Proportion of contaminated sites by industry division

Industry division	Number of contaminated sites	Number of uncontaminated sites	Proportion of contaminated sites (%)
Oil mining (OM)	82	112	42.27
Ferrous metal ores mining and dressing (FMD)	1,367	962	58.69
Non-ferrous metal mining and dressing (NFMD)	1,876	1,163	61.73
Textile (TX)	2,568	6,375	28.72
Tanning of leather and fur (TLF)	1,531	1,143	57.26
Papermaking (PM)	207	125	62.35
Chemical (CM)	16,649	14,322	53.76
Chemical APIs (CAPIS)	1,777	1,069	62.44
Ferrous metal smelting (FMS)	3,237	2,111	60.53
Non-ferrous metal smelting (NFMS)	3,427	2,732	55.64
Electroplating (ET)	6,207	4,981	55.48
Refined petroleum products (RPP)	1,734	1,434	54.73
Coking (CK)	746	604	55.26
Battery (BT)	1,697	1,126	60.11
Waste disposal (WD)	571	1,563	26.76
Total	43,676	39,822	52.31

sites are mainly located in the eastern region, including Jiangsu, Zhejiang, Guangdong, and Shandong provinces, while the number of contaminated sites in western China is relatively small (Figure 2). To explain this difference with statistical data, the eastern part of China has 61.61% of the country's contaminated sites, while the western part has only 13.12% (Table S3). 80% of the eastern provinces have more contaminated sites than Sichuan province, which has the largest number of contaminated sites in western China (Table S3). The distribution pattern of contaminated sites is similar to China's economic development level, which decreases from the coast to the interior. However, according to the distribution density of contaminated sites in the urban area, the high density of contaminated sites is mainly concentrated in the southern regions of China, such as the provinces of Yunnan, Hunan, Guangxi, and Guizhou.

While contaminated sites are widely distributed in China's urban areas, they are mostly concentrated in a few regions, showing remarkable regional variation. As shown in Figure 3B, contaminated sites in seven developed urban agglomerations account for 53.41% of the country.[37,38] The urban agglomeration with the highest number and density of contaminated sites is the Yangtze River Delta urban agglomeration (YRD) with 11,916 sites and 0.48 sites per square kilometer. The number and density of contaminated sites show large spatial variations. The degree of urbanization of the agglomeration is positively reflected in the number of contaminated sites, while the density is not. For example, the number of contaminated sites in the Pearl River Delta urban agglomeration (PRD) is 4.29 times higher than in the Chang-Zhu-Tan (CZT), while they have the same density of contaminated sites in the urban area (Figure 3B). This indicates that the number of contaminated sites is more consistent with the level of urbanization than the density. Spatial variability exists not only between but also within urban agglomerations. Taking the Mid-southern Liaoning urban agglomeration (MSL) as an example, the number of contaminated sites in Shenyang is 6.94 times higher than that in Tieling, and the density of contaminated sites in Benxi is 5.13 times higher than that in Dalian (Figure 3F).

Among the industry divisions of contaminated sites, CM, ET, and NFMS account for the largest number of three industrial types, covering 60.10% of the country (Figure 4A). Driven by resource endowment, labor force, market, related industries, and development history, the industry divisions of contaminated sites in the urban area also show remarkable spatial variation. For example, affected by both labor force and domestic market, contaminated sites in CM, ET, and textile (TX), are the main polluting industries in the YRD and PRD, accounting for 77.5% and 72.2%, respectively (Figures 4B and 4C). The rich non-ferrous metal mining resources in the Central Yunnan urban agglomeration (CY) make the NFMD the main polluting industry (Figure 4G).

Temporal changes of contaminated sites

Massive urban expansion has led to a 3.25 times increase in contaminated sites from 1990 to 2018 (Table S3), but the urban area has expanded 6.55 times. The urban area was divided into four urban sub-areas (core urban, sub-core urban, new urban, and suburbs) which were built in different periods (before 1990, 1991–2000, 2001–2010, and 2011–2018) to analyze the temporal changes of contaminated sites in the urban area. The density of contaminated sites was used to represent changes between the four urban sub-areas. As can be seen from Figure 5A, from a national perspective, the density of contaminated sites decreases from the core urban area to the suburbs and shows significant differences ($p < 0.05$) between the four urban sub-areas. However, not all urban agglomerations exhibit this significant difference due to their history of industrial development and urban expansion. For example, with the exception of the PRD, the density of contaminated sites in the
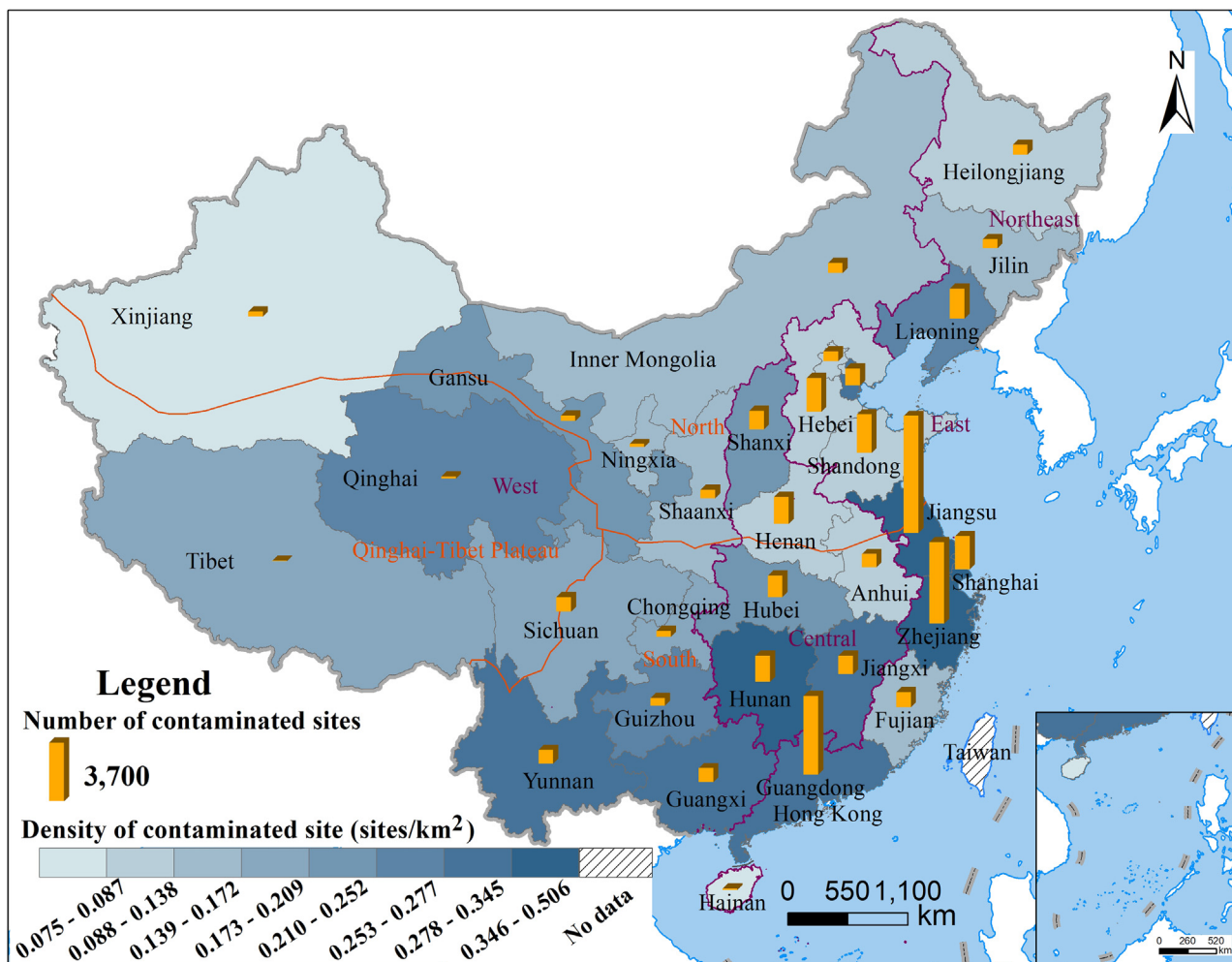
Figure 2. Spatial pattern of contaminated sites in China

The darker map color indicates a higher density of contaminated sites in the urban area of the province. The height of the column indicates the number of contaminated sites in the urban area of the province.

core urban area of the other urban agglomerations is not significantly different from that in the sub-core urban area, indicating that urban expansion is accompanied by the development of polluting enterprises during this period.

Compared to the general distribution of the contaminated sites in the four urban sub-areas, there are some differences by industry divisions. For example, resource-dependent industries, such as oil mining, ferrous metal mining and dressing, ferrous metal smelting, and cooking, which promoted the emergence and development of cities, tend to be distributed in the core urban area (Figure S1). Most of TX and ET are located in the new urban area, where these enterprises can benefit from the labor force, energy, and upstream enterprises (Figure S1).

As shown in Figures 6 and 7, the status and changes in the number of polluting enterprises causing site pollution in the four urban sub-areas vary considerably between 1990 and 2020. The number of new polluting enterprises has experienced three periods, namely accelerated growth, decelerated growth, and slow growth (Figure 6A). Before 1998, during the period of accelerated growth, the growing number of new polluting enterprises in the four urban sub-areas decreased from the inside out (Core > Sub-core > New > Suburbs). While the growth of polluting enterprises in the new urban area and suburbs reached its maximum between 1998 and 2008, the four urban sub-areas generally showed a decelerated growth trend, especially the core urban area. Since 2008, the number of new polluting enterprises has entered a period of slow growth. Although the overall pattern of change in new polluting enterprises in each urban agglomeration is similar to that of the country, there are some differences in the four urban sub-areas due to their different industrial development histories. In the YRD and CY, most of the new polluting enterprises were established in the new urban area, while in PRD they were in the sub-core area, and in MSL, Beijing-Tianjin-Hebei urban agglomeration (BTH), and Chengdu-Chongqing urban agglomeration (CC) they were in the core urban area (Figures 6B, 6C, and 6E–6H).

Overall, the number of shutdown enterprises continues to grow (Figure 7A). The cumulative number of shutdown enterprises has risen from 1,326 to 13,233 over the last two decades. The number of shutdown polluting enterprises that have closed is highest in the core area, while the
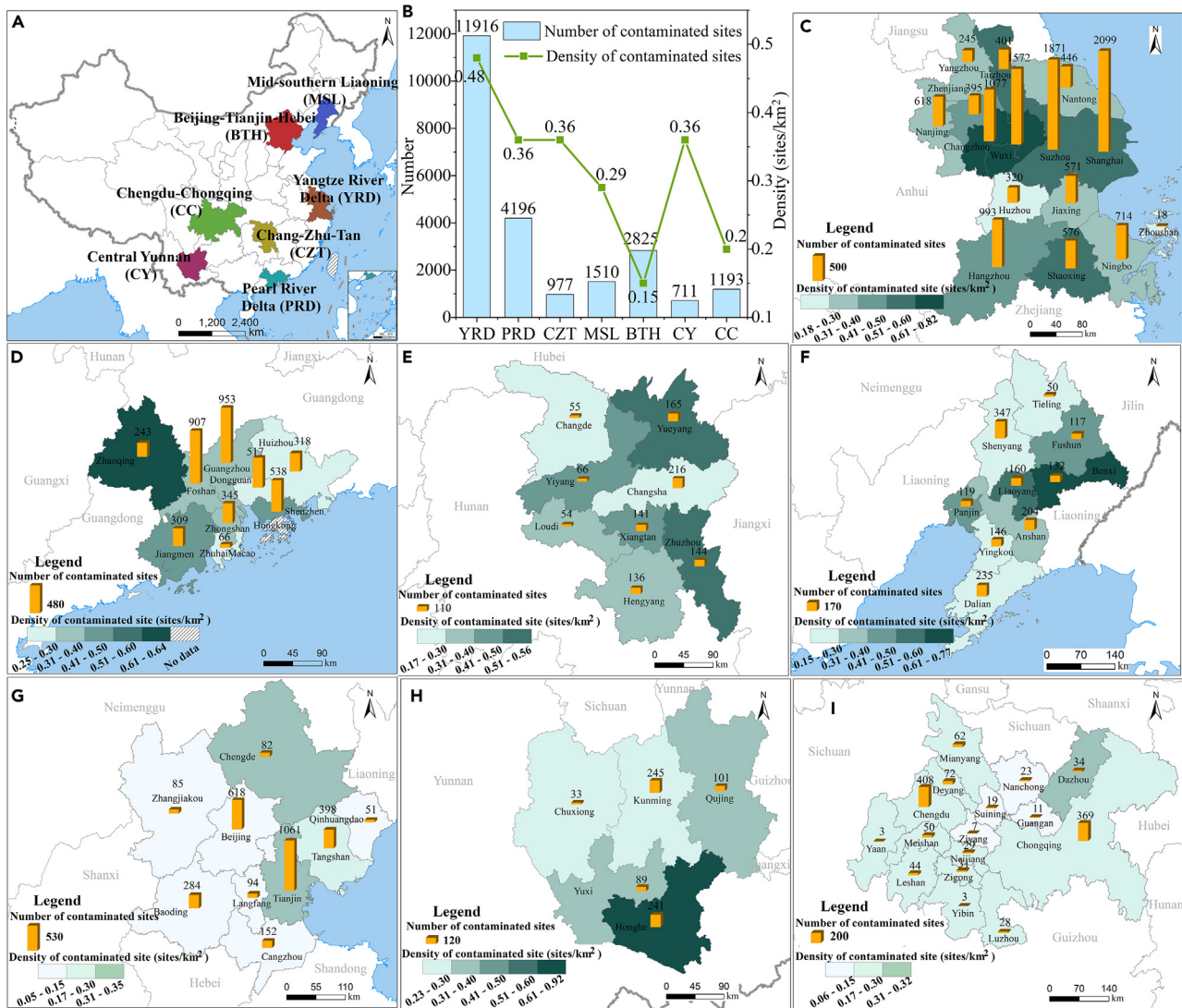
Figure 3. Spatial pattern of contaminated sites in urban agglomerations

The darker map color indicates a higher density of contaminated sites in the urban area of the city. The height of the column indicates the number of contaminated sites in the urban area of the city.

(A) Spatial distribution of major urban agglomerations in China.

(B) The left and right y axis show the number and density of contaminated sites in the urban area of the seven urban agglomerations, respectively.

(C) YRD = Yangtze River Delta urban agglomeration.

(D) PRD = Pearl River Delta urban agglomeration.

(E) CZT = Chang-Zhu-Tan urban agglomeration.

(F) MSL = Mid-southern Liaoning urban agglomeration.

(G) BTH = Beijing-Tianjin-Hebei urban agglomeration.

(H) CY = Central Yunnan urban agglomeration.

(I) CC = Chengdu-Chongqing urban agglomeration.

rate of shutdown polluting enterprises in the new urban area has accelerated in recent years and may exceed that of the core area next year (Figure 7A). The pattern of change in the number of shutdown polluting enterprises between urban sub-areas of agglomerations is similar to that of new polluting enterprises. These changing trends of fewer new polluting enterprises and more shutdown polluting enterprises have led to a decrease in the number of contaminated sites in production since 2008.

Urban development has led to the relocation or decommissioning of polluting industries to the periphery of the city. As can be seen from Figures 6A and 7A, after 1998, the new urban areas overtook the core urban areas to become the urban areas with the highest number of new polluting enterprises. Meanwhile, the number of closed polluting enterprises in the core urban area remained the highest and showed an
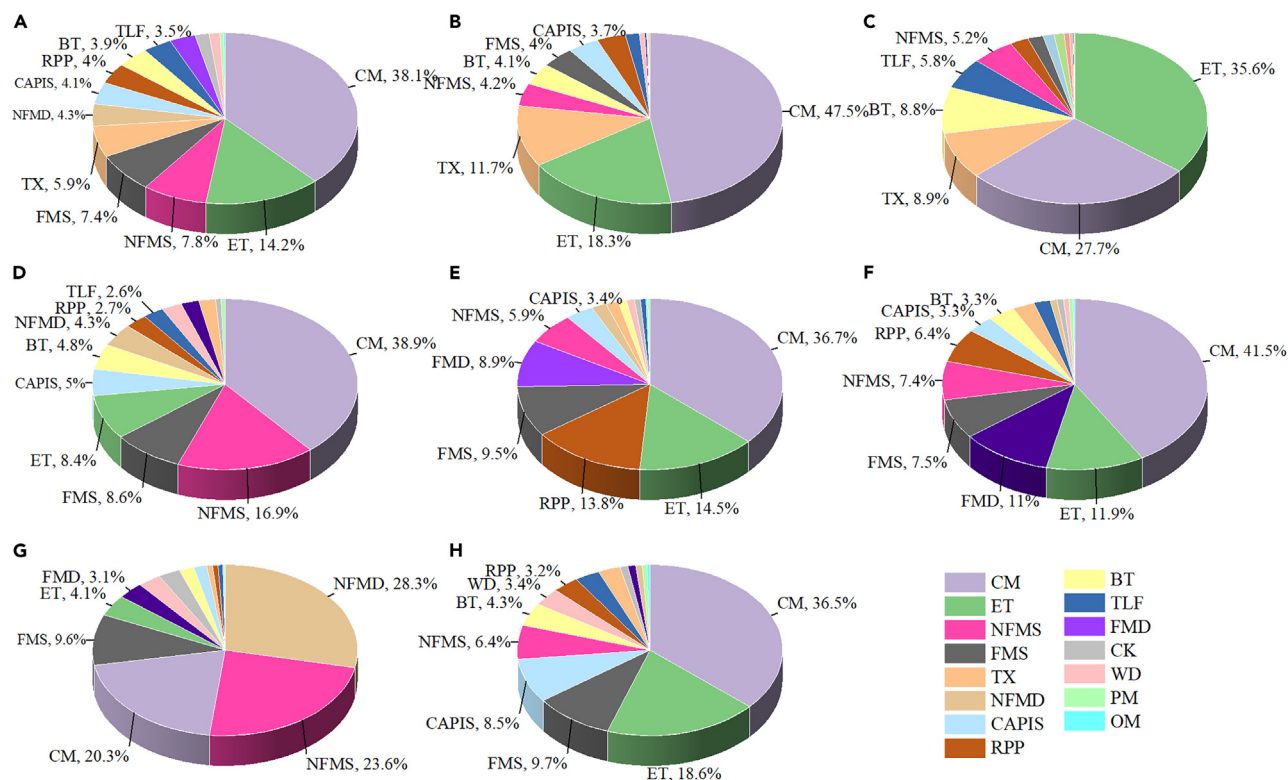
Figure 4. Percentage of contaminated sites by industry division
(A) Country.
(B) YRD.
(C) PRD.
(D) CZT.
(E) MSL.
(F) BTH.
(G) CY.
(H) CC.

accelerated growth. This indicates an accelerated transfer of contaminated sites from the core urban area to the new urban area. The variations in the number of active polluting enterprises between different urban sub-areas also confirmed this phenomenon (Figure S2A). Overall, there is an initial increase and then a decrease in the number of active polluting enterprises in each urban sub-area. However, the rate of growth is particularly rapid in the new urban areas, which overtook the core urban areas by 2005. From the perspective of agglomerations, the rate at which polluting industries relocate to the periphery of cities is slower in areas with earlier economic development. For example, the BTH and MSL are the earliest developed industrial bases in China, and as a result consistently have the highest number of active polluting enterprises in their core urban areas (Figures S2E and S2F). In contrast, the YRD has consistently had the highest number of active polluting enterprises in its new urban areas (Figure S2B).

DISCUSSION

BLR model uncertainty

The uncertainty of the regression coefficients of the BLR model is presented in Table S4. All of the coefficient estimates have p values less than 0.05, indicating that they are statistically significant. The t-values are relatively large for all variables, further supporting the significance of the coefficient estimates. In addition, the standard errors are relatively small, suggesting that the coefficient estimates are reliable. The 95% confidence intervals provide a range of plausible values for the coefficients. The narrower the confidence interval, the more precise the estimation of the coefficient. In this case, the confidence intervals for all variables are relatively narrow, indicating a higher level of precision in the estimation. Overall, the results suggest that the model coefficients are statistically significant and have reliable estimates.

The average AUC value across the 5-folds is 0.86, indicating a good overall performance of the model. However, it is also important to assess the variability or uncertainty of the AUC values. The AUC values obtained from cross-validation provide an estimate of the model's performance on unseen data. A larger standard deviation implies that the model's performance is less consistent across different samples,
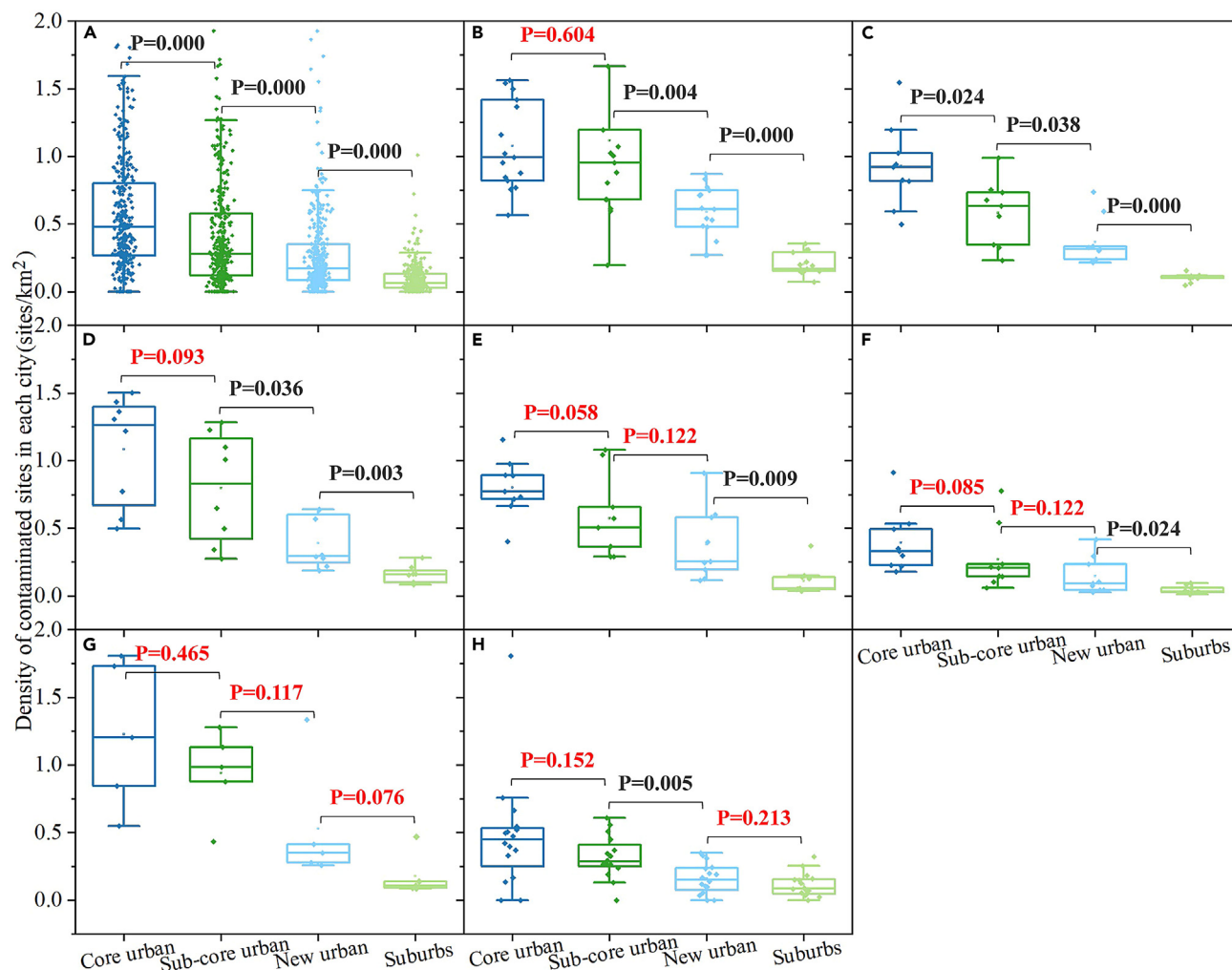
Figure 5. Spatial variation of contaminated sites between the four urban sub-areas

The density of contaminated sites in the four urban sub-areas was plotted using boxplots and tested using the Mann-Whitney U test, with red markers indicating no significant difference and black markers indicating a significant difference (p < 0.05).

(A) Country.
(B) YRD.
(C) PRD.
(D) CZT.
(E) MSL.
(F) BTH.
(G) CY.
(H) CC.

indicating higher uncertainty. On the other hand, a smaller standard deviation suggests more consistent performance and lower uncertainty. The standard deviation of the AUC values is 0.03, suggesting a low level of variability in the model's performance across different folds. This indicates that the performance of the model is consistent and stable.

Potential drivers of spatiotemporal variation of contaminated sites

Although the causes of the spatiotemporal variation of contaminated sites in the urban area are very complex, such differences are largely a result of economic development level, resource endowment, urban expansion, production and operation status of the enterprise, and natural conditions. Based on the results of our research, a discussion is presented on the drivers of spatiotemporal variation of polluting enterprises and the main control factors of site pollution in the urban area.

The economic development level has a positive feedback relationship with the cumulative number of enterprises, especially enterprises with high pollution and energy consumption.[39,40] On the one hand, the economically developed regions provide convenient transportation,
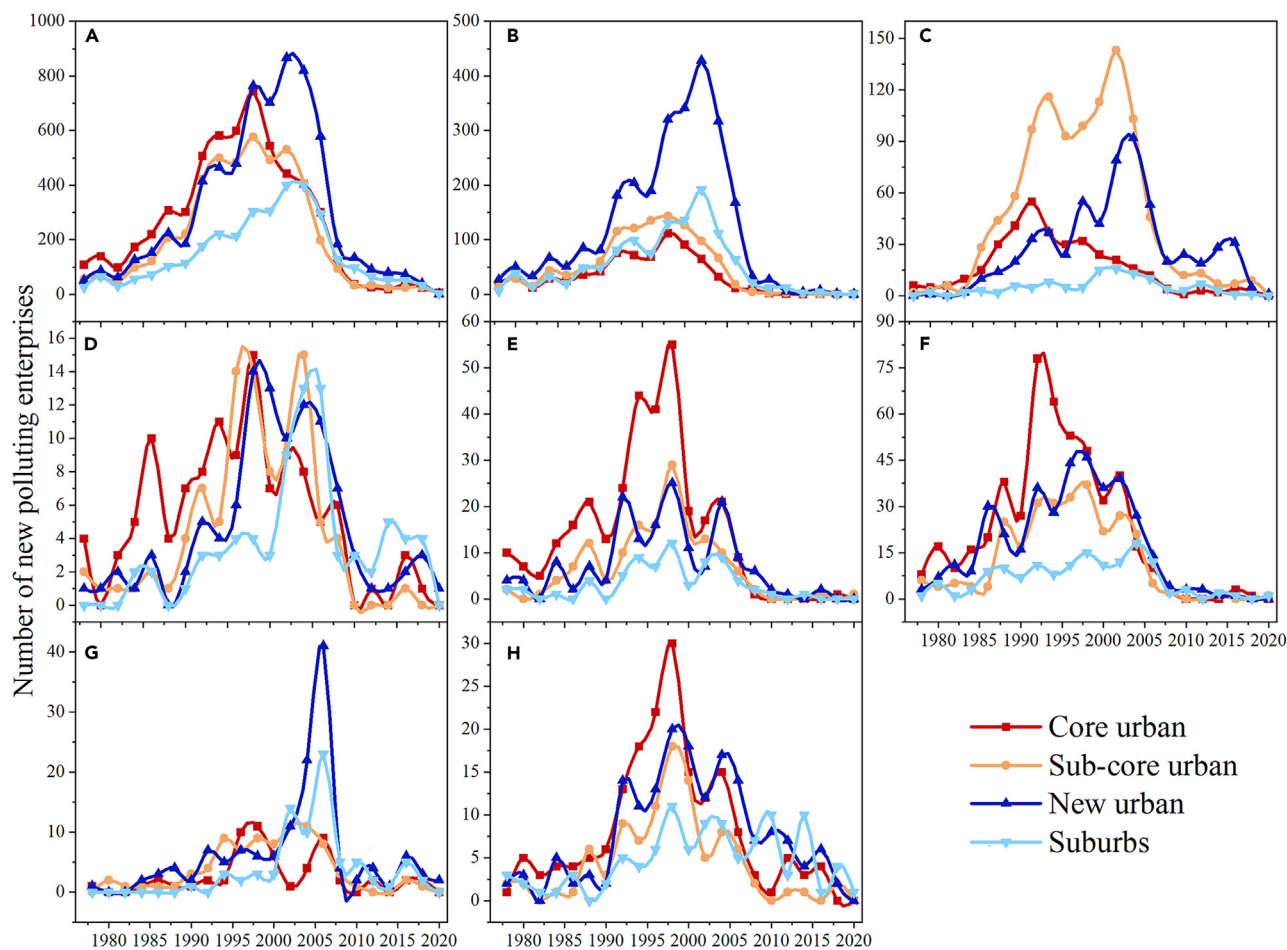
Figure 6. The number change of new polluting enterprises causing site pollution in four urban sub-areas from 1978 to 2020
(A) Country.
(B) YRD.
(C) PRD.
(D) CZT.
(E) MSL.
(F) BTH.
(G) CY.
(H) CC.

complete production facilities, sufficient energy, and a broad market for the development of polluting enterprises. On the other hand, polluting enterprises, as low-end manufacturing industries, are the main types of enterprises that promote economic development and urbanization in regions and countries at the early stage of industrialization.[41] As shown in Figures 5A and 6A, the construction of polluting enterprises is earlier than that of urban areas. Different from the developed regions, resource endowment is the main driving factor for the development of polluting enterprises in underdeveloped regions.[42] For example, 51.43% of the contaminated sites in underdeveloped western China are related to mineral-resource-processing enterprises, while developed eastern China only accounts for 13.30%.

Due to the diverse driving mechanisms,[43] the seven urban clusters showed differences in the spatiotemporal distribution characteristics of contaminated sites. Because the YRD and PRD are the most developed region in China, it has become the area with the highest concentration of contaminated sites.[44,45] Convenient transportation, proximity to consumer markets, well-established industrial chains, abundant labor force, and reform and opening-up policy have been significant factors contributing to the aggregation of polluting enterprises in the CM, ET, and TX industries.[33,46] The well-developed water and rail transport systems, together with the comprehensive infrastructure, and proximity to the YRD, provide favorable conditions for the development of manufacturing and CM industries in the CZT. In addition, the abundance of mineral resources is the main reason for the concentration of non-ferrous metal and steel enterprises in the region.[47,48] The BTH, being a political, cultural, and early industrialization center in China, has attracted a significant number of manufacturing enterprises.[49] Abundant resources of oil, non-ferrous metals, and iron ore in the BTH require the establishment of numerous extraction and processing companies.[50]
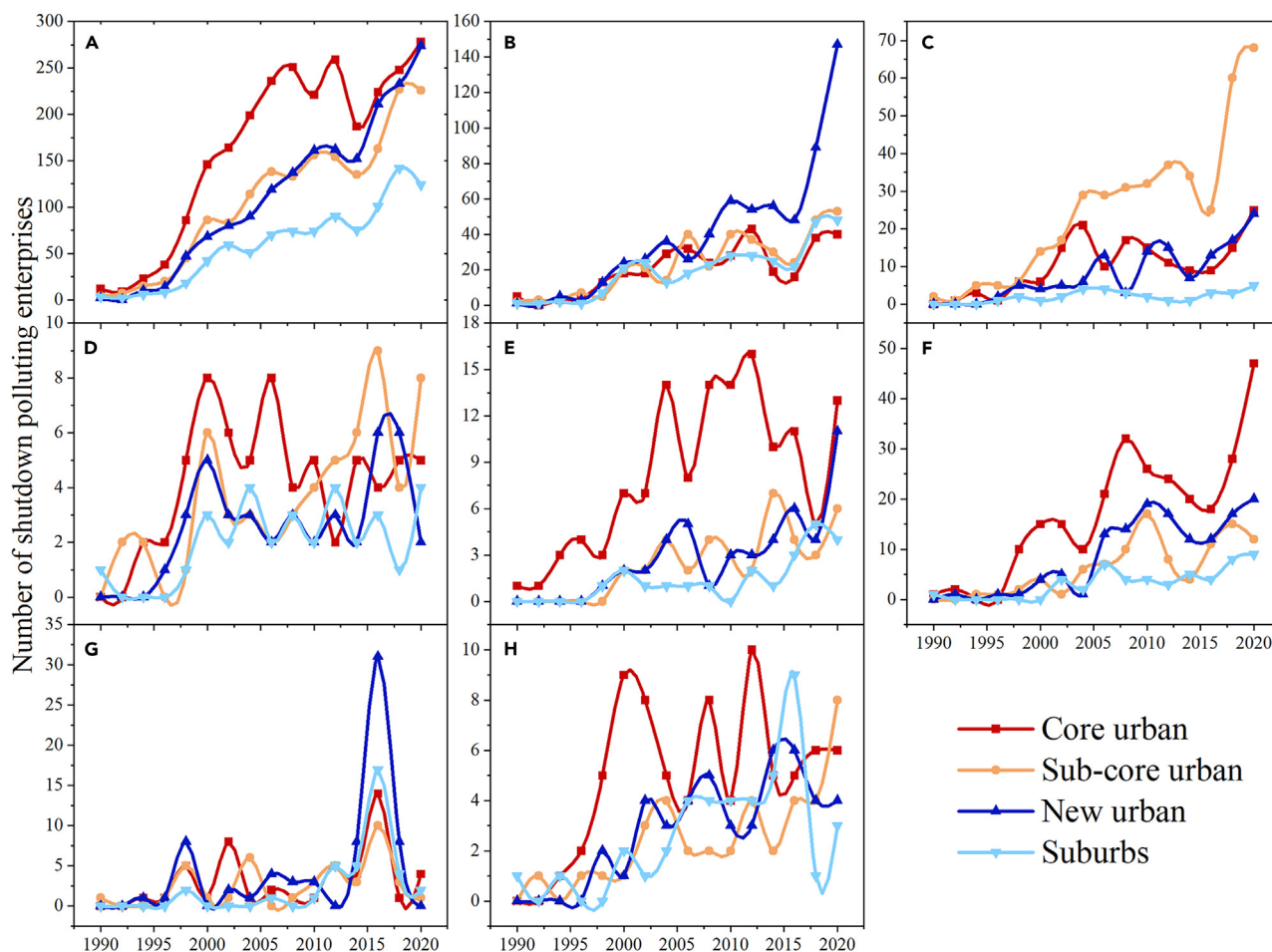
Figure 7. The number change of shutdown polluting enterprises causing site pollution in four urban sub-areas from 1990 to 2020
(A) Country.
(B) YRD.
(C) PRD.
(D) CZT.
(E) MSL.
(F) BTH.
(G) CY.
(H) CC.

Similar to the BTH, the MSL serves as one of China's historical industrial production bases and harbors abundant resources of oil, non-ferrous metals, and iron ore.[51,52] The CY, known as the "Non-Ferrous Kingdom," has extensive contaminated sites, mainly due to the mining and smelting of non-ferrous metals in the region.[42] The CC serves as a manufacturing hub in western China, acting as an agglomeration area for the production of chemicals, non-ferrous metals, pesticides, and steel.[53]

The spatial variation of polluting enterprises is the basis for the spatial variation of contaminated sites. According to our research, whether they become contaminated sites is related to the production and operation status of the enterprise and the natural conditions. As shown in Figure 1C, duration, starting time, industry class, violations, precipitation, and temperature are the main drivers of site pollution. The longer the duration of industrial production, the more likely it is that the pollutants will accumulate on the enterprise plot by running, emitting, dripping, and leaking.[54] Polluting enterprises established in different periods have used different processes, and emissions have decreased as science and technology have progressed.[55] The raw materials, auxiliary materials, processes, products, and pollutant types vary widely between enterprises in different industry classes.[56–58] The number of enterprise environmental violations, such as the illegal discharge of pollutants and environmental accidents, directly reflects the enterprise's environmental management level. As a result of erosion and rainfall run-off, pollutants in raw materials, wastewater, and solid waste enter the soil and groundwater through horizontal migration and infiltration.[59] Through volatilization to the atmosphere and migration to the deeper soil, the temperature has a significant impact on the diffusion of soil organic pollutants.[60,61]

Implications for urban environmental management

In this study, the trained BLR model, using only 6 publicly available variables, showed high accuracy in identifying contaminated sites (Figure 1A). These results support that the machine-learning method applied to identify contaminated sites is fast, efficient, and accurate, as demonstrated by previous pollution researchers.[33,62,63] Before the Chinese government has enough money and time to investigate, evaluate, and remediate the polluting enterprise plots, machine learning is a useful method to rank the polluting enterprise plots and give priority to the control and management of enterprise plots with a high pollution probability. Moreover, advances in indicator selection and ranking methods using machine learning make it possible to simplify the index system and reduce the cost of the environmental risk assessment of polluting enterprise plots used by the Chinese government.[64] As can be seen from Figure S3, the kernel density analysis was used to rank the spatial density of contaminated sites in the urban area. The hot zones with kernel density values higher than 0.01 sites per square kilometer are consistent with the spatial distribution of heavy-metal-emitting enterprises[43] and contaminated sites investigated in China (Figure S4). This result has potential applications in demarcating the key areas for soil environmental management and industrial upgrading.

In addition to the spatial variation across China, our analysis shows that contaminated sites in the urban area have characteristics of temporal change, which provides governors with some new ideas to improve the management efficiency of contaminated sites. For example, the number of contaminated sites opened from 2011 to 2020 accounts for only 3.5% of the total number of contaminated sites, while the number of contaminated sites opened from 1991 to 2000 accounts for 41.71% (Figure S5). The proportion of polluting sites opened between 2011 and 2020 is only 22.89%, compared with 66.66% between 1991 and 2000 (Figure S5). The longer a polluting enterprise has been in operation, the more likely it is that its land will be contaminated (Figure 1C). When the operational duration of a polluting enterprise exceeds 15 years, the probability of its land being contaminated increases rapidly (Figure S5). Due to the significant differences in the management strategies and pollution characteristics between shutdown and in production polluting enterprises,[31] more attention should be paid to the significant changes over time in these enterprises in the four urban sub-areas. First, the environmental protection and industrial upgrading policies in the past decade have turned rapid growth into low growth of contaminated sites, and this growth trend will continue. Second, the rapid growth of the shutdown enterprises will bring great pressure to the investigation, risk assessment and control, restoration, and reuse of the contaminated sites, especially in the core and new urban areas (Figure 7A). Third, the number of polluting enterprises in production with contaminated sites will continue to decrease in the future through merging multiple enterprises and improving the production process (Figure S2). Therefore, different soil environment strategies should be adopted based on the production duration and status of polluting enterprises to improve efficiency.

Limitations of the study

Because of limited data availability, we collected information on 83,498 polluting enterprises in the urban area, which are the soil pollution industries that managers focus on regulating. Under these conditions, the 43,676 contaminated sites identified by the machine-learning model are likely to be smaller than the actual number. Although the trained BLR model achieved high accuracy by learning knowledge from 2,005 samples, the accuracy of the model will improve if more samples are obtained from non-public government data. Furthermore, four commonly used machine-learning models were employed for contaminated site identification. Future research could explore the utilization of alternative models to further improve the model accuracy and apply the model to more polluting industry divisions.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Data collection and processing
  - Methods

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.108124.

AUTHOR CONTRIBUTIONS

Conceptualization, K.L. and R.S.; methodology, K.L.; resources, K.L.; formal analysis, K.L. and G.G.; visualization, K.L.; writing – original draft, K.L.; writing – review & editing, R.S. and G.G.; supervision, R.S.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Zhou, W., Yu, W., Qian, Y., Han, L., Pickett, S.T.A., Wang, J., Li, W., and Ouyang, Z. (2022). Beyond city expansion: multi-scale environmental impacts of urban megaregion formation in China. Natl. Sci. Rev. 9, 107. https://doi.org/10.1093/nsr/nwab107.

2. Li, X., Gong, P., Zhou, Y., Wang, J., Bai, Y., Chen, B., Hu, T., Xiao, Y., Xu, B., Yang, J., et al. (2020). Mapping global urban boundaries from the global artificial impervious area (GAIA) data. Environ. Res. Lett. 15, 094044. https://doi.org/10.1088/1748-9326/ab9be3.

3. Gong, P., Li, X., Wang, J., Bai, Y., Chen, B., Hu, T., Liu, X., Xu, B., Yang, J., Zhang, W., and Zhou, Y. (2020). Annual maps of global artificial impervious area (GAIA) between 1985 and 2018. Remote Sens. Environ. 236, 111510. https://doi.org/10.1016/j.rse.2019.111510.

4. Hou, W., Zhang, L., Li, Y., Zhang, L., Li, S., Ji, L., and Dan, S. (2015). Distribution and Health Risk Assessment of Polycyclic Aromatic Hydrocarbons in Soil from a Typical Contaminated Urban Coking Sites in Shenyang City. B. Environ. Contam. Tox. 95, 815–821. https://doi.org/10.1007/s00128-015-1677-3.

5. Yang, H., Flower, R.J., and Thompson, J.R. (2012). Rural factories won't fix Chinese pollution. Nature 490, 342–343. https://doi.org/10.1038/490342d.

6. MNR (Ministry of Natural Resources of the People's Republic of China); MEE (Ministry of Ecology and Environment of the People's Republic of China) (2014). National Soil Pollution Survey Bulletin. https://www.mee.gov.cn/gkml/sthjbgw/qt/201404/W020140417558995804588.pdf.

7. Fang, Y., Nie, Z., Die, Q., Tian, Y., Liu, F., He, J., and Huang, Q. (2017). Organochlorine pesticides in soil, air, and vegetation at and around a contaminated site in southwestern China: Concentration, transmission, and risk evaluation. Chemosphere 178, 340–349. https://doi.org/10.1016/j.chemosphere.2017.02.151.

8. Neupane, A., and Gustavson, K. (2008). Urban property values and contaminated sites: A hedonic analysis of Sydney, Nova Scotia. J. Environ. Manag. 88, 1212–1220. https://doi.org/10.1016/j.jenvman.2007.06.006.

9. Yang, H., Huang, X., Thompson, J.R., and Flower, R.J. (2014). China's Soil Pollution: Urban Brownfields. Science 344, 691–692. https://doi.org/10.1126/science.344.6185.691-b.

10. Bai, X., Chen, J., and Shi, P. (2012). Landscape Urbanization and Economic Growth in China: Positive Feedbacks and Sustainability Dilemmas. Environ. Sci. Technol. 46, 132–139. https://doi.org/10.1021/es202329f.

11. Grimm, N.B., Faeth, S.H., Golubiewski, N.E., Redman, C.L., Wu, J., Bai, X., and Briggs, J.M. (2008). Global change and the ecology of cities. Science 319, 756–760. https://doi.org/10.1126/science.1150195.

12. Shochat, E., Lerman, S.B., Anderies, J.M., Warren, P.S., Faeth, S.H., and Nilon, C.H. (2010). Invasion, Competition, and Biodiversity Loss in Urban Ecosystems. Bioscience 60, 199–208. https://doi.org/10.1525/bio.2010.60.3.6.

13. Foley, J.A., Defries, R., Asner, G.P., Barford, C., Bonan, G., Carpenter, S.R., Chapin, F.S., Coe, M.T., Daily, G.C., Gibbs, H.K., et al. (2005). Global consequences of land use. Science 309, 570–574. https://doi.org/10.1126/science.1111772.

14. Liu, J., and Diamond, J. (2005). China's environment in a globalizing world. Nature 435, 1179–1186. https://doi.org/10.1038/4351179a.

15. Deng, X., Huang, J., Rozelle, S., Zhang, J., and Li, Z. (2015). Impact of urbanization on cultivated land changes in China. Land Use Policy. 45, 1–7. https://doi.org/10.1016/j.landusepol.2015.01.007.

16. Zhou, D., Zhao, S., Liu, S., Zhang, L., and Zhu, C. (2014). Surface urban heat island in China's 32 major cities: Spatial patterns and drivers. Remote Sens. Environ. 152, 51–61. https://doi.org/10.1016/j.rse.2014.05.017.

17. Kalnay, E., and Cai, M. (2003). Impact of urbanization and land-use change on climate. Nature 423, 528–531. https://doi.org/10.1038/nature01675.

18. Zhou, L., Dickinson, R.E., Tian, Y., Fang, J., Li, Q., Kaufmann, R.K., Tucker, C.J., and Myneni, R.B. (2004). Evidence for a significant urbanization effect on climate in China. Proc. Natl. Acad. Sci. USA 101, 9540–9544. https://doi.org/10.1073/pnas.0400357101.

19. Hu, X., Zhou, W., Qian, Y., and Yu, W. (2017). Urban expansion and local land-cover change both significantly contribute to urban warming, but their relative importance changes over time. Landsc. Ecol. 32, 763–780. https://doi.org/10.1007/s10980-016-0484-5.

20. Han, L., Zhou, W., Li, W., and Li, L. (2014). Impact of urbanization level on urban air quality: A case of fine particles (PM2.5) in Chinese cities. Environ. Pollut. 194, 163–170. https://doi.org/10.1016/j.envpol.2014.07.022.

21. Miller, J.D., and Hutchins, M. (2017). The impacts of urbanisation and climate change on urban flooding and urban water quality: A review of the evidence concerning the United Kingdom. J. Hydrol-Reg. Stud. 12, 345–362. https://doi.org/10.1016/j.ejrh.2017.06.006.

22. Li, X., Jiao, W., Xiao, R., Chen, W., and Liu, W. (2017). Contaminated sites in China: Countermeasures of provincial governments. J. Clean. Prod. 147, 485–496. https://doi.org/10.1016/j.jclepro.2017.01.107.

23. Li, X.N., Jiao, W.T., Xiao, R.B., Chen, W.P., and Chang, A.C. (2015). Soil pollution and site remediation policies in China. Environ. Rev. 23, 263–274. https://doi.org/10.1139/er-2014-0073.

24. Zhang, R., Jiang, L., Zhong, M., Han, D., Zheng, R., Fu, Q., Zhou, Y., and Ma, J. (2019). Applicability of Soil Concentration for VOC-Contaminated Site Assessments Explored Using Field Data from the Beijing-Tianjin-Hebei Urban Agglomeration. Environ. Sci. Technol. 53, 789–797. https://doi.org/10.1021/acs.est.8b03241.

25. Wcisło, E., Bronder, J., Bubak, A., Rodríguez-Valdés, E., and Gallego, J.L.R. (2016). Human health risk assessment in restoring safe and productive use of abandoned contaminated sites. Environ. Int. 94, 436–448. https://doi.org/10.1016/j.envint.2016.05.028.

26. Kuppusamy, S., Palanisami, T., Megharaj, M., Venkateswarlu, K., and Naidu, R. (2016). In-Situ Remediation Approaches for the Management of Contaminated Sites: A Comprehensive Overview. Rev. Environ. Contam. Toxicol. 236, 1–115. https://doi.org/10.1007/978-3-319-20013-2_1.

27. Mahammedi, C., Mahdjoubi, L., Booth, C.A., Akram, H., and Butt, T.E. (2020). A systematic review of risk assessment tools for contaminated sites - Current perspectives and future prospects. Environ. Res. 191, 110180. https://doi.org/10.1016/j.envres.2020.110180.

28. Mahammedi, C., Mahdjoubi, L., Booth, C.A., and Butt, T.E. (2022). Framework for preliminary risk assessment of brownfield sites. Sci. Total Environ. 807, 151069. https://doi.org/10.1016/j.scitotenv.2021.151069.

29. Kovalick, W.W., and Montgomery, R.H. (2017). Models and lessons for developing a contaminated site program: An international review. Environ. Technol. Innov. 7, 77–86. https://doi.org/10.1016/j.eti.2016.12.005.

30. USEPA (U.S. Environmental Protection Agency) (1990). Hazard Ranking System. https://semspub.epa.gov/work/HQ/174028.pdf.

31. MEE (Ministry of Ecology and Environment of the People's Republic of China) (2017). Notice on Printing and Distributing a Series of Technical Documents on Land Use Survey of Enterprises in Key Industries. https://www.mee.gov.cn/gkml/hbb/bgt/201708/t20170818_420021.htm.

32. Fathizad, H., Ardakani, M.A.H., Heung, B., Sodaiezadeh, H., Rahmani, A., Fathabadi, A., Scholten, T., and Taghizadeh-Mehrjardi, R. (2020). Spatio-temporal dynamic of soil quality in the central Iranian desert modeled with machine learning and digital soil assessment techniques. Ecol. Indicat. 118, 106736. https://doi.org/10.1016/j.ecolind.2020.106736.

33. Jia, X., Hu, B., Marchant, B.P., Zhou, L., Shi, Z., and Zhu, Y. (2019). A methodological framework for identifying potential sources of soil heavy metal pollution based on machine learning: A case study in the Yangtze Delta,

China. Environ. Pollut. *250*, 601–609. https://doi.org/10.1016/j.envpol.2019.04.047.

34. Jia, Z., Zhou, S., Su, Q., Yi, H., and Wang, J. (2017). Comparison Study on the Estimation of the Spatial Distribution of Regional Soil Metal(loid)s Pollution Based on Kriging Interpolation and BP Neural Network. Int. J. Environ. Res. Publ. Health *15*, 34. https://doi.org/10.3390/ijerph15010034.

35. Podgorski, J.E., Labhasetwar, P., Saha, D., and Berg, M. (2018). Prediction Modeling and Mapping of Groundwater Fluoride Contamination throughout India. Environ. Sci. Technol. *52*, 9889–9898. https://doi.org/10.1021/acs.est.8b01679.

36. Christian, F.S. (2015). SPSS Regression Analysis (Publishing House of Electronics Industry).

37. Fang, C.L. (2018). Important Progress and Prospects of China's Urbanization and Urban Agglomeration in the Past 40 Years of Reform and Opening-Up. Econ. Geogr. *38*, 1–9. https://doi.org/10.15957/j.cnki.jjdl.2018.09.001.

38. Li, J.M., Zhang, W.Z., Sun, T.S., and Zhang, A.P. (2014). Characteristics of clustering and economic performance of urban agglomerations in China. Acta Geograph. Sin. *69*, 474–484. https://doi.org/10.11821/dlxb201404004.

39. Liu, W., Shen, J., Wei, Y.D., and Chen, W. (2021). Environmental justice perspective on the distribution and determinants of polluting enterprises in Guangdong, China. J. Clean. Prod. *317*, 128334. https://doi.org/10.1016/j.jclepro.2021.128334.

40. Wang, Y., Duan, X., and Wang, L. (2020). Spatial distribution and source analysis of heavy metals in soils influenced by industrial enterprise distribution: Case study in Jiangsu Province. Sci. Total Environ. *710*, 134953. https://doi.org/10.1016/j.scitotenv.2019.134953.

41. Li, X., Zhang, X., Du, H., and Chu, S. (2012). Spatial effect of mineral resources exploitation on urbanization: A case study of Tarim River Basin, Xinjiang, China. Chin. Geogr. Sci. *22*, 590–601. https://doi.org/10.1007/s11769-012-0554-9.

42. Wu, Y., Li, X., Yu, L., Wang, T., Wang, J., and Liu, T. (2022). Review of soil heavy metal pollution in China: Spatial distribution, primary sources, and remediation alternatives. Resour. Conserv. Recycl. *181*, 106261. https://doi.org/10.1016/j.resconrec.2022.106261.

43. Li, K., Wang, J., and Zhang, Y. (2022). Heavy metal pollution risk of cultivated land from industrial production in China: Spatial pattern and its enlightenment. Sci. Total Environ. *828*, 154382. https://doi.org/10.1016/j.scitotenv.2022.154382.

44. Fu, X., Wang, S., Zhao, B., Xing, J., Cheng, Z., Liu, H., and Hao, J. (2013). Emission inventory of primary pollutants and chemical speciation in 2010 for the Yangtze River Delta region, China. Atmos. Environ. *70*, 39–50. https://doi.org/10.1016/j.atmosenv.2012.12.034.

45. Wong, S.C., Li, X.D., Zhang, G., Qi, S.H., and Min, Y.S. (2002). Heavy metals in agricultural soils of the Pearl River Delta, South China. Environ. Pollut. *119*, 33–44. https://doi.org/10.1016/S0269-7491(01)00325-6.

46. Zhao, W., Ma, J., Liu, Q., Dou, L., Qu, Y., Shi, H., Sun, Y., Chen, H., Tian, Y., and Wu, F. (2023). Accurate Prediction of Soil Heavy Metal Pollution Using an Improved Machine Learning Method: A Case Study in the Pearl River Delta, China. Environ. Sci. Technol. https://doi.org/10.1021/acs.est.2c07561.

47. Liu, H., Probst, A., and Liao, B. (2005). Metal contamination of soils and crops affected by the Chenzhou lead/zinc mine spill (Hunan, China). Sci. Total Environ. *339*, 153–166. https://doi.org/10.1016/j.scitotenv.2004.07.030.

48. Hu, B., Shao, S., Ni, H., Fu, Z., Hu, L., Zhou, Y., Min, X., She, S., Chen, S., Huang, M., et al. (2020). Current status, spatial features, health risks, and potential driving factors of soil heavy metal pollution in China at province level. Environ. Pollut. *266*, 114961. https://doi.org/10.1016/j.envpol.2020.114961.

49. Liang, L., Wang, Z., and Li, J. (2019). The effect of urbanization on environmental pollution in rapidly developing urban agglomerations. J. Clean. Prod. *237*, 117649. https://doi.org/10.1016/j.jclepro.2019.117649.

50. Yu, C., Li, H., Jia, X., and Li, Q. (2015). Improving resource utilization efficiency in China's mineral resource-based cities: A case study of Chengde, Hebei province. Resour. Conserv. Recycl. *94*, 1–10. https://doi.org/10.1016/j.resconrec.2014.10.003.

51. Qing, X., Yutong, Z., and Shenggao, L. (2015). Assessment of heavy metal pollution and human health risk in urban soils of steel industrial city (Anshan), Liaoning, Northeast China. Ecotoxicol. Environ. Saf. *120*, 377–385. https://doi.org/10.1016/j.ecoenv.2015.06.019.

52. Liu, S.S., Li, H.Y., Du, X.M., Wang, H.B., Lv, C.Y., Wang, X., and Li, F.S. (2007). Variation of Industrial Pollution Sources and Driving Factor Analysis in Liaoning Province. Res. Environ. Sci. *20*, 142–148. https://doi.org/10.13198/j.res.2007.06.145.liushsh.018.

53. Liu, Q., Li, X., and He, L. (2022). Health risk assessment of heavy metals in soils and food crops from a coexist area of heavily industrialized and intensively cropping in the Chengdu Plain, Sichuan, China. Front. Chem. *10*, 988587. https://doi.org/10.3389/fchem.2022.988587.

54. Wang, Q., Hao, D., Wang, F., Wang, H., Huang, X., Li, F., Li, C., and Yu, H. (2021). Development of a new framework to estimate the environmental risk of heavy metal(loid)s focusing on the spatial heterogeneity of the industrial layout. Environ. Int. *147*, 106315. https://doi.org/10.1016/j.envint.2020.106315.

55. Li, X., Xiao, R., Chen, W., Chang, C., Deng, Y., and Xie, T. (2017). A Conceptual Framework for Classification Management of Contaminated Sites in Guangzhou, China.

Sustainability *9*, 362. https://doi.org/10.3390/su9030362.

56. Li, C. (2017). Study on Emission Inventory of Arsenic Pollution Sources in Copper Smelting Process (Beijing University of Chemical Technology).

57. Li, J. (2015). Typical Heavy Metal Pollution Identification, Prevention and Control Countermeasures in Nickel and Cobalt Smelting (China University of Mining&Technology).

58. Nriagu, J.O., and Pacyna, J.M. (1988). Quantitative Assessment of Worldwide Contamination of Air, Water and Soils by Trace-Metals. Nature *333*, 134–139. https://doi.org/10.1038/333134a0.

59. Mikelonis, A.M., Hawley, R.J., and Goodrich, J.A. (2021). Emergency response to stormwater contamination: A framework for containment and treatment. J. Environ. Manag. *280*, 111838. https://doi.org/10.1016/j.jenvman.2020.111838.

60. Jones, D.A., Lelyveld, T.P., Mavrofidis, S.D., Kingman, S.W., and Miles, N.J. (2002). Microwave heating applications in environmental engineering—a review. Resour. Conserv. Recycl. *34*, 75–90. https://doi.org/10.1016/S0921-3449(01)00088-X.

61. Zhang, H., Yuan, X., Xiong, T., Wang, H., and Jiang, L. (2020). Bioremediation of co-contaminated soil with heavy metals and pesticides: Influence factors, mechanisms and evaluation methods. Chem. Eng. J. *398*, 125657. https://doi.org/10.1016/j.cej.2020.125657.

62. Bešter, P.K., Lobnik, F., Eržen, I., Kastelec, D., and Zupan, M. (2013). Prediction of cadmium concentration in selected home-produced vegetables. Ecotoxicol. Environ. Saf. *96*, 182–190. https://doi.org/10.1016/j.ecoenv.2013.06.011.

63. Novotná, M., Mikeš, O., and Komprdová, K. (2015). Development and comparison of regression models for the uptake of metals into various field crops. Environ. Pollut. *207*, 357–364. https://doi.org/10.1016/j.envpol.2015.09.043.

64. MEE (Ministry of Ecology and Environment of the People's Republic of China) (2017). Technical Regulations on Land Risk Screening and Risk Grading of Enterprises in Production. https://www.mee.gov.cn/gkml/hbb/bgt/201708/W020170818581370130828.pdf.

65. MEE (Ministry of Ecology and Environment of the People's Republic of China) (2018). Soil Environmental Quality Risk Control Standard for Soil Contamination of Development Land. https://www.mee.gov.cn/ywgz/fgbz/bz/bzwb/trhj/201807/t20180703_446027.shtml.

66. Overmars, K.P., and Verburg, P.H. (2005). Analysis of land use drivers at the watershed and household level: Linking two paradigms at the Philippine forest fringe. Int. J. Geogr. Inf. Sci. *19*, 125–152. https://doi.org/10.1080/13658810410001713380.

67. Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recogn. Lett. *27*, 861–874. https://doi.org/10.1016/j.patrec.2005.10.010.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Sample data set of surveyed sites	Official government websites, Commercial websites	https://doi.org/10.17632/r4y2vcpfmx.1
Simulation data set	Official government websites, Commercial websites	https://doi.org/10.17632/r4y2vcpfmx.1
Software and algorithms		
BLR code	Scikit-learn	https://doi.org/10.17632/r4y2vcpfmx.1
RF code	Scikit-learn	https://doi.org/10.17632/r4y2vcpfmx.1
BP code	Scikit-learn	https://doi.org/10.17632/r4y2vcpfmx.1
SVM code	Scikit-learn	https://doi.org/10.17632/r4y2vcpfmx.1
ArcGIS	ESRI	https://www.arcgis.com/index.html
Origin 2021	OriginLab	https://www.originlab.com/OriginProLearning.aspx
Anaconda	Continuum Analytics	https://www.anaconda.com/

RESOURCE AVAILABILITY

Lead contact
Further information for data and code files should be directed to and will be fulfilled by the lead contact, Ranhao Sun (rhsun@rcees.ac.cn).

Materials availability
This study did not generate new unique reagents.

Data and code availability
DATA: The input dataset and simulation dataset have been deposited at Mendeley Data and are publicly available as of the date of publication. The DOI is listed in the key resources table.
CODE: All original code used for constructing the machine-learning models has been deposited at Mendeley Data and is publicly available as of the date of publication. The DOI is listed in the key resources table.

METHOD DETAILS

Data collection and processing

Sample of surveyed sites

As an input sample for the machine-learning model, the 2,005 surveyed sites were collected from public information including repaired plots, risk control plots, surveyed plots, and studied sites from literature (Figure S6A). As shown in Tables S5 and S6, the data sources for this information are official government websites, Green Data, Web of Science, and CNKI using web crawling and manual searches. If the contamination levels of the surveyed sites exceed the screening values for Class I land use,[65] they are classified as contaminated sites.

Simulation data set

The simulation data set of the machine-learning model is 83,498 polluting enterprise plots, consisting of pollutant discharge permit enterprises, backward capacity enterprises, heavy metal emission enterprises in key industries, key supervision enterprises of soil environmental pollution, and key pollutant discharge enterprises (Figure S6B). The official government websites are the main source of these data (Tables S5 and S6).

Variable preparation for machine-learning modeling

Fourteen available variables from public information were selected as potential variables in the statistical modeling based on established or presumed relationships with site pollution (Table S8). However, owing to the diversity of sample and simulation data, we supplemented the missing attribute information by the open network to obtain information on these variables, such as the National Enterprise Credit Information Publicity System, Institute of Public and Environmental Affairs, Tianyancha, and Green Data (Table S5), and then cleaned the incomplete

and repeated data. The detailed situation of missing data and the method of data completion are shown in Table S7. The complete attributes of the sample and simulation data, including name, status, starting time, closed time, registered capital, industry class, number of environmental violations, and production address, were obtained, and the sample data additionally contained the type and content of pollutants. Furthermore, the geographical coordinates of the polluting enterprises in the sample and simulation data were translated using geographical coding technology based on the production address attribute. The values of the natural variables were extracted from the associated GIS data sets (Table S8) at each polluting enterprise. The detailed acquisition method and type of potential variables is shown in Table S8. As shown in Table S9, the categorical variables are transformed into numerical variables to allow the model to read the variables.

Urban expansion data

A 30m resolution global urban boundary dataset (http://data.ess.tsinghua.edu.cn/) in seven representative years (i.e. 1990, 1995, 2000, 2005, 2010, 2015, and 2018) was used as urban expansion data in China. This dataset was generated by an automatic delineation framework using 30 m global artificial impervious area (GAIA) data.[2] The urban boundaries are consistent well with results derived from nighttime light data and human interpretation.

Methods

T-test and chi-square test

To eliminate spurious variables, reduce noise and training time, and avoid falling into local minima in the machine-learning process, the appropriate variable screening methods of t-test and chi-square test were used to obtain the minimum number of independent variables to ensure the model accuracy. The t-test was used for continuous variables and the chi-square test was used for categorical variables.

Binary logistic regression

Discrimination of contaminated sites is a binary classification problem in machine-learning classification Compared with linear regression and logarithmic regression, BLR is an algorithm specially used for the discontinuous or categorical dependent variable.[66] The core algorithms of BLR are the sigmoid function, maximum likelihood estimation, loss function, and gradient descent algorithm. We supposed that the variables x_1, x_2, x_3, x_4.....of the investigated site are a set of variables related to the dependent variable Y, and the value of Y is 1 (contaminated site) or 0 (uncontaminated site). The regression model is as follows:

$$P(Y_i = 1) = \frac{\exp\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ji}\right)}{1 + \exp\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ji}\right)}$$

(Equation 1)

$$\text{logit}P(Y_i = 1) = \log\left(\frac{P_i}{1 - P_i}\right) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ji} + e$$

(Equation 2)

$P_i = P(Y_i = 1)$: the probability of site i pollution; k: the number of variables; β_0: the constant term; β_j: the regression coefficient corresponding to explanatory variable x_j; e: the error term.

The ROC curve and ACC were used to evaluate the model performance. The ROC curve was plotted with sensitivity (TPR) as the ordinate and 1-specificity (FPR) as the abscissa to produce the associated area under the ROC curve (AUC) value, which generally ranges between 0.5 (no predictive capability) and 1 (perfect predictive capability).[67] TPR refers to the proportion of correct positive samples in all positive samples and represents the discriminant ability of the model to positive samples. FPR is the proportion of negative samples with wrong judgment in all negative samples. The highest value of the Youden index is the threshold to balance sensitivity and specificity. ACC refers to the proportion of correct judgments in all samples, which can be used to evaluate the overall accuracy rate. K-fold cross-validation was used to obtain k times performance evaluation indicators to verify the accuracy and stability of model results.