Front. Eng. Manag. 2022, 9(1): 163–169
https://doi.org/10.1007/s42524-021-0185-x

COMMENTS

Miaomiao LIU, Bing ZHANG, Jun BI

Appreciating the role of big data in the modernization of environmental governance

1 Introduction

The development of the Internet of Things, drones, and new social media platforms opened up an unprecedented space for the autonomy of human society. It promoted the formation of a new form of social relations and social field — cyber society (Jones, 1995). The cyber society has already become a digital twin of the actual society and even the extensions. It reconstructed the whole society through impacting social control and social norms. In this context, it is of great theoretical and practical significance to expand the field of social governance from traditional society to cyber society.

As a vital information carrier, data records all activities in reality and cyber society. At present, cheaper and more accessible data storage, analysis, sharing, and distribution tools brought an explosive and exponential growth of the global data volume. Both governments and business companies have realized the massive value of big data. They took their respective advantages to dig out information for business management and social governance decision-makings from massive and unstructured big data. With artificial intelligence technologies such as machine learning, image identification, and natural language processing, we achieved significant advancements in data collection, processing, optimization, and prediction. These advancements allow us to observe the social operation mechanisms from big data that help the decision-makers improve the modernization level of social governance. In this context, the role of big data in social governance has become a hot topic in the current management research.

As an essential field of social governance, modernizing the systems and capacities of environmental governance is an important dimension. Meanwhile, the field of environmental governance offers a broader space for applying big data and artificial intelligence due to its characteristics of multiple spatial-temporal scales, complex interaction effects between pollutants and polluting media, and high uncertainties involving stakeholders (Zhong et al., 2021). A systemic understanding of the role of big data and artificial intelligence in promoting the transformation of environmental governance is in line with national strategic needs and has cross-field demonstration significance. Up to now, most research on big data in environmental governance has been case studies, explaining the technical paths of particular application scenarios. They revealed the partial picture but failed to form systemic thinking.

This paper puts forward a conceptual framework for the role of big data in environmental governance. The frontier application scenarios that deserve attention are summarized to elaborate the framework. It also highlights research frontiers of interest to scholars from a wide range of disciplines. Finally, from the viewpoint of system engineering, this paper specifies strategies that should be taken to cope with the challenges in China's practices.

2 Conceptual framework

This section proposes a conceptual framework of the role of big data in environmental governance (Fig. 1). It can be described as a demand–supply matching process. The combinations of environmental big-data and artificial intelligence tools provide multi-dimensional solutions to meet the needs of modernized environmental governance.

2.1 Environmental governance demands

On the demand side, modernized environmental governance seeks accurate, fast, and low-cost solutions to the

Received November 13, 2021; accepted December 8, 2021

Miaomiao LIU, Bing ZHANG (✉), Jun BI
State Key Laboratory of Pollution Control and Resource Reuse, School of Environment, Nanjing University, Nanjing 210023, China
E-mail: zhangb@nju.edu.cn

This work is supported by the National Natural Science Foundation of China (Grant Nos. 71921003, 72174084, and 72161147002) and the Fundamental Research Funds for the Central Universities (Grant No. 0211-14380171).

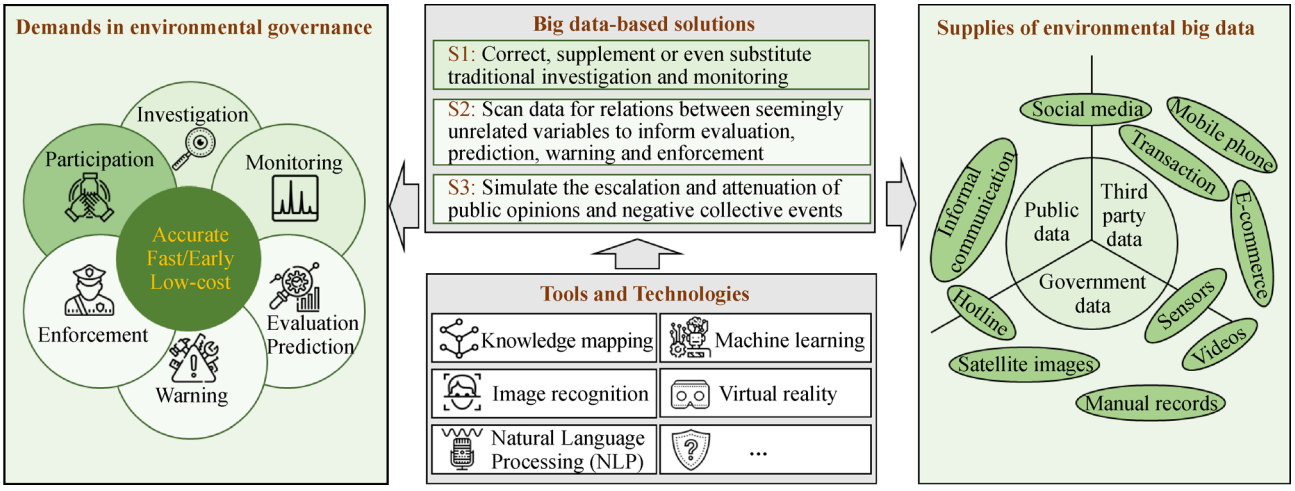

Fig. 1 Conceptual framework of the role of big data in environmental governance.

environmental information collection, evaluation and prediction, early warning and enforcement, and public supervision. This subsection elaborates on the urgent demands of current and future environmental governance.

2.1.1 Demands for low-cost and high-resolution environmental information

Environmental investigation and monitoring are two virtual channels to collect environmental information, serving as the basis of environmental science and practices. In recent years, China has carried out several large-scale special investigations nationwide, including the "2nd round of pollution sources census", "illegal solid waste sites investigations" in 2018, "biodiversity survey", and "urban black-odorous water body investigation". Meanwhile, China's monitoring networks on environmental quality, pollution sources, and ecology are rapidly developing from the perspectives of coverage, resolution, and diversity.

Despite those progresses, there are still many challenges. On the one hand, the current large-scale investigations are cross-sectional surveys at the cost of time and professionals. Therefore, they only offer static information about the environment and ecology status, sometimes even out-of-date information. On the other hand, the existing monitoring networks have low space–time resolutions and limited capacity to capture characteristic pollutants, ecological indicators, and climate hazards. Therefore, reducing the cost of environmental investigations, improving the timeliness of the investigation data, and supplementing the indicators of characteristic pollutants, ecological indicators, and climate disaster indicators into monitoring networks are crucial demands of improving the information collection capacity of the environmental governance system.

2.1.2 Demands for scientific and efficient environmental evaluations and predictions

Environmental evaluations and predictions are important bridges to transform the collected information into informative regulation decisions. However, China's current environmental evaluation practices highly depend on experts' judgment, requiring considerable resources and causing potential disputes. Taking an environmental risk evaluation practice in Jiangsu Province as an example, they organized around 500 experienced experts and 40 senior officials to evaluate the safety and environmental risks of 2600 chemical enterprises after several months of on-site surveys in 2019. Based on experts' evaluations, the officers decided whether to shut down, rectify, relocate, or upgrade each enterprise. However, the expert-based decisions attracted extensive concerns from enterprises.

In addition, the predictions of environmental trends mainly depend on physical or chemistry model simulations. For example, the global climate model, regional air quality model, and watershed hydrodynamic model have been widely used to predict the global temperature rise, $PM_{2.5}$ concentration, and water quality under different scenarios and further guide the formulation or selection of regulation strategies (Wang et al., 2021; Zhu et al., 2021). However, the inherent uncertainties in these models mean that new independent methods are required to verify and support the predictions. In this context, improving the scientific level and efficiency of environmental evaluations and predictions has become a significant demand.

2.1.3 Demands for early warning and targeted enforcement

Apart from predicting macro eco-environmental trends, it is also vital to warn and regulate illegal behaviors of individual enterprises, such as illegal discharge dumping.

At present, we have collected massive behavioral data of individual enterprises by video, sensors, drones, online monitoring equipment, manual sampling, etc. But the data are interpreted to be early-warning information for illegal behaviors mainly through manual or simple logical judgments. For example, the management personnel detects the illegal discharge behaviors by artificially observing the colors of wastewater in the videos or setting the fixed pollutant concentration thresholds for online monitoring equipment. These practices face the problems of high manual dependence, strong randomness, and poor warning effects. Similar paradoxes challenge the regulations of waste gas and hazardous waste.

Moreover, China implemented double-random environmental law enforcement, randomly selecting inspected objects and inspectors. As a result, it may allocate limited law enforcement resources to objects with low violation probabilities, resulting in low efficiencies. Simultaneously, the potential selection bias leads to the misjudges of decision-makers on the local environmental performance. In this context, it is urgent to investigate the enterprise's illegal environmental motivation from massive environmental data, informing effective early warning and law enforcement.

2.1.4 Demands for rational public participation and supervision

In addition to governments and enterprises, participation and supervision from the public are also important for environmental governance. Unfortunately, the high-frequency and unordered interactions of social media platforms make the public unable to distinguish biased opinions. In the face of sudden interventions, they are prone to anxiety and fear, resulting in the amplification of short-term social risks. Taking the public opinions after an intervention of the *Under the Dome* video as an example, the information delivered by the video magnified the air pollution risk perceived by the public in the short term. It also presented long-term and intensified effects of undermining public trust in local governments (Liu et al., 2021). Moreover, the real-time information dissemination and exchange in the cyber society accelerates the transformation from biased public to irrational participation behaviors and even negative collective events, threatening the stability of society. Therefore, guiding rational public participation and supervision is also an important demand.

2.2 Big data-based solution supply

The combinations of big data and artificial intelligence tools are expected to provide multi-dimensional solutions for the abovementioned demands. This subsection summarizes the sources of big data of the environment and promising solutions.

In the widely-used 5V model, big data has five characteristic dimensions of high volume, variety, velocity, value, and veracity. Although the data involved in environmental science and practice cannot fully meet the strict definition of big data in terms of volume dimension, its granularity, complexity, real-time ability, and accuracy have gradually possessed the basic characteristics of big data. According to broad definitions, environmental big data comes from a wide range of sources, including the governments, third parties, and the public.

Government data are usually from the large-scale official statistics and surveys, satellite remote sensing, high-altitude radar, drone observations, ground monitoring station, video monitoring, sensors, and online monitoring. Another important data source is the video or sensor data collected by third parties such as non-governmental organizations on the public welfare ground. It shall be highlighted that big data of the environment is not limited to the data directly relevant to environmental media or pollution sources. Substantial seemingly unrelated data can also serve environmental science and practices in particular scenarios. Such data includes mobile signaling, social media, bank transactions, electricity consumption, and e-commerce. In addition, the public also generates useful data, of which the most typical example is public environmental complaints. With the development of mobile terminal equipment and We-media platforms, the public plays an increasingly important role in environmental data collection and supply.

Combining environmental big-data and artificial intelligence tools can provide solutions from the following dimensions. Firstly, the assimilation and integration of big data from satellite remote sensing, video, GPS, and online monitoring equipment improve data collection capacities through verifying the data reliability, filling in the blank, and refining the space-time resolution. Moreover, all subjects in the cyber society generate and supply data, promoting the information collection modes shifted from top-down ones dominated by the governments to bottom-up ones with whole-society participation. Massive third-party and social media information can supplement or even substitute traditional labor-intensive and time-consuming information collection channels. Secondly, big data, together with artificial intelligence tools, is an ideal fit for environmental evaluation, prediction, early warning, and regulation because it is excellent at scanning substantial data sets for patterns and relationships between seemingly unrelated variables. Unfortunately, these relationships may have been covered up in traditional data analysis, hindering smart decisions. Thirdly, the explosive social media data and the development of community discovery algorithms and natural language processing tools enable us to simulate the escalation and attenuation of public opinions and even negative collective events. It identifies the timing of risk communications and improves adaptability and recoverability after the shocks. Notable

application scenarios in frontier research are described in Section 3 to help readers understand these solutions.

3 Notable application scenarios in frontier research

Around the abovementioned environmental governance needs, international scholars have launched innovative exploration. This section describes the frontier application scenarios worthy of attention from multiple disciplines.

3.1 Environmental information collection scenarios

The most common and mature applications of big data in environmental governance are using machine learning models to assimilate multi-source heterogeneous investigation and monitoring data to improve spatial-temporal coverage and resolution. In particular, international scholars pay unprecedented attention to simulating high-resolution concentration distributions of air pollutants such as NO_2, $PM_{2.5,}$ and O_3 (ozone) using the data from satellite remote sensing observations, atmospheric chemical transmission models, and surface monitoring stations (Reid et al., 2015; Di et al., 2016). For example, Di et al. (2016) used a backward propagation neural network to calibrate GEOS-Chem simulations and predicted $PM_{2.5}$ concentrations in 1 km $\times$ 1 km grid cells in the Northeastern United States. Liu et al. (2020) used daily maximum 8-hour average (MDA8) ozone observations from 2013 to 2017 combined with concurrent ozone retrievals, aerosol reanalysis, meteorological parameters, and land-use data to establish a nationwide MDA8 prediction model based on the eXtreme Gradient Boosting algorithm. They further simulated the ozone MDA8 concentration distributions in China from 2005 to 2019, filling the blank of ground monitoring data before 2013.

International researchers gradually expand this kind of application from air pollution to more indicators. For example, Wen et al. (2018) provided a solution for investigating and monitoring urban black-odor water bodies based on the Gaofen-2 (GF-2) satellite image data modeling. Lombard et al. (2021) developed a machine learning model using enhanced regression tree and random forest classification (RFC) technology to estimate probabilities and concentration ranges of arsenic in private wells throughout the conterminous United States at a resolution of 1 km. Yang et al. (2021a) developed a machine learning model for soil adsorption of six heavy metals (Cd (II), Cr (VI), Cu (II), Pb (II), Ni (II), and Zn (II)) using 4420 data points (1105 soils) extracted from 150 journal articles. They simulated the heavy metal adsorption capacity of soils around the world. Moore and Obradovich (2020) found that compared with the extrapolations from a sparse network of tide gauges, the number of flood-related tweets that naturally integrates a measure of the social consequences of flooding can better estimate localized (i.e., county-specific) flooding thresholds.

Upon the latest researches, two clear trends have been developed. Firstly, big data's applications expand from common pollutants to characteristic pollutants, urban black-odor water bodies, and climate hazards, for which data were more challenging to collect and monitor (Zhong et al., 2021). The second trend is the enrichment of data sources. Besides traditional data sources like satellite remote sensing, ground monitoring, and chemical transmission mode, data from lab tests, academic journals, news reports, and social media serve as supplements and even replacements to government-dominated environment investigation and monitoring data (Moore and Obradovich, 2020; Yang et al., 2021a). These trends enlighten directions for the big data future applications within environment information collection.

3.2 Environmental evaluation, prediction, early-warning, and enforcement scenarios

Unlike the application in information collection, big data in environmental evaluation, prediction, early warning, and regulation scenarios are fragmented.

Some are relevant to studies that predict and regulate eco-environmental quality. They focus on integrating big data and machine learning to verify the traditional evaluation and prediction results from the global climate model, regional air quality model, and watershed hydrodynamic model, and consequently reduce the uncertainty of policy intervention. For example, based on integrated meteorological observation, surface air quality monitoring, and high-precision emission data, Vu et al. (2019) applied a random forest technique to predict Beijing's action plan's effectiveness after decoupling the impact of meteorology on ambient air quality. They provided a solution to eliminating the inherent uncertainty in the traditional air quality model. Based on real-time road traffic activity data, meteorological observation, points of interest, and other data, Yang et al. (2021b) used a random forest model to predict the impact of COVID-19 lockdown and electrification policy on urban NO_2 concentration in Los Angeles. It provided big data-based evidence for electrification policy selection.

Researchers also find interesting application scenarios in environmental behavior early warning. For example, Lu (2019) identified illegal dumping cases by mining a publicly available data set containing more than 9 million waste disposal records from 2011 to 2017. He identified possible causes for illegal dumping (e.g., long queuing times) and produced a list of 546 waste hauling trucks suspected of involvement in illegal dumping using behavioral indicators and up-to-date big data analytics. In addition to illegal behavior warnings, researchers also used big data to regulate responsive behaviors. For example, Almuhtaram et al. (2021) proposed a novel machine

learning model to determine when a utility needs to respond to a harmful algal bloom by identifying anomalies in phycocyanin fluorescence data from online monitoring probes. Eyre et al. (2020) predicted the post-emergency recovery of small businesses after natural hazard events by analyzing their online posting activity data on social media, guiding the allocation of government emergency resources.

Referring to the latest studies and the abovementioned environment governance demands, we would like to highlight the potentials of big data and machine learning in early warning of illegal environment behaviors. For instance, to address illegal dumping of hazardous wastes, on the one hand, a machine learning model can be built to predict theoretical hazardous wastes production volume of enterprises based on their financial reports, raw material balance, electricity and production process monitor, online waste gas and water monitor, and self-reported hazardous waste data. The probability of illegally dumping hazardous wastes can be estimated by identifying the gap between theoretical and self-reported volume. On the other hand, the probability of different spatial locations becoming illegal dumping sites can be estimated based on the data from satellite remote sensing, drone monitor and local road network. Then, by overlapping the suspected illegal dumping sites with enterprises with a high probability of illegal dumping, the illegal dumping sources can be traced and early-alarmed. Furthermore, the machine-learning-based illegal behavior probability prediction model can greatly enhance the accuracy of law enforcement. Clearly, the application scenarios are not limited to these two cases. We expect more exploration and researches will bring this field to the next level.

3.3 Public participation and supervision

With the rise of the Internet of Things, drones, and new social media platforms, the channels for the public to obtain and disseminate information have become convenient. It alleviates information asymmetry faced by the public in traditional environmental governance and improves the importance of public participation in environmental governance. Big data, especially unstructured text data from social media, has been widely used in tracking public concerns and attitudes towards hot environmental issues. For example, according to Twitter data, Kirilenko and Stepchenkova (2014) interpreted the public's attitude towards climate change. They identified the users and organizations that the public believes are the most authoritative to provide climate change information.

In addition, studies also used big data to measure the public's response to environmental policies. For example, Wu et al. (2021) used text mining technology to measure the public's emotion and behavior towards the municipal solid waste classification policy based on Sina Weibo data. Similar data and research paradigms are also applied to measure the public's response to green buildings and prefabricated buildings (Liu and Hu, 2019).

The latest research reveals the unreplaceable advantages of big data in tracking hot environment issues, guiding public environmental opinions, and communicating with the public. Integrating big data from social media with natural language processing technologies to quickly identify hot environmental issues, critical regions, and key population groups and resolve issues in time is worthy of attention in future research.

4 Challenges and solutions in China's practices

China has put a heavy weight on the potential contributions of big data applications in environmental governance. As early as 2015, the General Office of the State Council of the PRC issued the Ecological Environment Monitoring Network Construction Plan. It pointed out that China shall build the ecological environment big data platform to provide data support for environmental protection. In 2016, the Ministry of Environmental Protection officially issued the Overall Plan for Big Data Construction of Ecological Environment, which further clarified the goals of big data applications in environmental fields. Nowadays, big data has served China's environmental quality forecast and early warning system, hazardous waste life-cycle tracing project, intelligent management of industrial parks, and many other practices. However, there are still many challenges in furthering big data application in China's environmental governance. This section summarizes the potential challenges and proposes specific solutions.

Firstly, "garbage in, garbage out (GIGO)" is the golden rule in data science. The subjects in many fields voluntarily and actively provide as real information as possible for personal benefits. For example, as the subjects in the health management field, the patients provide as real and accurate information as possible to help disease diagnosis and treatment. However, an important type of subject in the environmental field is enterprises. To avoid environmental responsibility, enterprises tend to avoid data collection conducted by the government or provide compliance data rather than real data. Thus, compared with other fields, the quality of big data in the environmental field is usually worse and faces a high risk of falsifying data. Currently, big data used in environmental governance practices comes from diverse sources and has complex structures that lack standardized quality assessment and control. Therefore, it is common to spend weeks or months on data cleaning before carrying out big data-based environmental research and practices. In the face of this great challenge, data sharing is more important in the environmental field than in other fields. The establishment of an open-access data-sharing community can increase the volume and

diversity of data by adding more data sources on the one hand. On the other hand, it can realize mutual verification of data quality from different sources. More importantly, in this way, researchers and policymakers can achieve data quality control by sharing the experience of using data and avoiding the poor decisions that result from noisy data.

Secondly, big data-based environmental governance usually requires the support of algorithms and tools. By introducing multi-disciplinary algorithms and tools such as artificial intelligence, researchers have successfully applied big data to environmental pollution and climate disaster prediction, early warning of illegal environmental behaviors, environmental public opinion tracking, and many other fields. In practice, big data-based environmental governance requires officials to be constantly aware of new, interdisciplinary algorithms and tools. Such continuous commitments required can be difficult. It, to some extent, hinders the deep applications of big data. The current practices of environmental governance based on big data are still in the stage of big data collection and visualization, transforming data to information supporting supervision and decision-making through manual judgment or simple logical judgment. To maximize the value of big data, it is necessary to strengthen the construction of think tanks, open up communication channels between the government and scientific institutions, and promote the fast transformation of scientific and technological achievements into local practice.

Thirdly, data security and privacy are the challenges that all data science and practices cannot ignore. Therefore, it is critical to carry out systemic research on relevant laws and regulations, institutional norms, and standard guidelines in the global community. We shall identify the regulatory gaps in China's current environmental big data application and develop a regulatory system for security and privacy that is suitable for China.

Last but not least, the application of big data may have long-term, deep and unpredictable impacts on the industrial division of labor, technological substitution, and social structure. Therefore, we encourage follow-up studies on the impacts of big data applications in environmental governance to prevent systemic risks that may arise from the unknown.

References

Almuhtaram H, Zamyadi A, Hofmann R (2021). Machine learning for anomaly detection in cyanobacterial fluorescence signals. Water Research, 197: 117073

Di Q, Kloog I, Koutrakis P, Lyapustin A, Wang Y, Schwartz J (2016). Assessing $PM_{2.5}$ exposures with high spatiotemporal resolution across the continental United States. Environmental Science & Technology, 50(9): 4712–4721

Eyre R, de Luca F, Simini F (2020). Social media usage reveals recovery of small businesses after natural hazard events. Nature Communications, 11(1): 1629

Jones S G (1995). CyberSociety: Computer-Mediated Communication and Community. Thousand Oaks, CA: Sage Publications

Kirilenko A P, Stepchenkova S O (2014). Public microblogging on climate change: One year of Twitter worldwide. Global Environmental Change, 26: 171–182

Liu M, Bi J, Yang J, Qu S, Wang J (2021). Social media never shake the role of trust building in relieving public risk perception. Journal of Cleaner Production, 282: 124442

Liu R, Ma Z, Liu Y, Shao Y, Zhao W, Bi J (2020). Spatiotemporal distributions of surface ozone levels in China from 2005 to 2017: A machine learning approach. Environment International, 142: 105823

Liu X, Hu W (2019). Attention and sentiment of Chinese public toward green buildings based on Sina Weibo. Sustainable Cities and Society, 44: 550–558

Lombard M A, Bryan M S, Jones D K, Bulka C, Bradley P M, Backer L C, Focazio M J, Silverman D T, Toccalino P, Argos M, Gribble M O, Ayotte J D (2021). Machine learning models of arsenic in private wells throughout the conterminous United States as a tool for exposure assessment in human health studies. Environmental Science & Technology, 55(8): 5012–5023

Lu W (2019). Big data analytics to identify illegal construction waste dumping: A Hong Kong study. Resources, Conservation and Recycling, 141: 264–272

Moore F C, Obradovich N (2020). Using remarkability to define coastal flooding thresholds. Nature Communications, 11(1): 530

Reid C E, Jerrett M, Petersen M L, Pfister G G, Morefield P E, Tager I B, Raffuse S M, Balmes J R (2015). Spatiotemporal prediction of fine particulate matter during the 2008 northern California wildfires using machine learning. Environmental Science & Technology, 49(6): 3887–3896

Vu T, Shi Z, Cheng J, Zhang Q, He K, Wang S, Harrison R M (2019). Assessing the impact of clean air action on air quality trends in Beijing using a machine learning technique. Atmospheric Chemistry and Physics, 19(17): 11303–11314

Wang Y, Hu J, Zhu J, Li J, Qin M, Liao H, Chen K, Wang M (2021). Health burden and economic impacts attributed to $PM_{2.5}$ and O_3 in China from 2010 to 2050 under different representative concentration pathway scenarios. Resources, Conservation and Recycling, 173(38): 105731

Wen S, Wang Q, Li Y M, Zhu L, Lü H, Lei S H, Ding X L, Miao S (2018). Remote sensing identification of urban black-odor water bodies based on high-resolution images: A case study in Nanjing. Environmental Science, 39(1): 57–67 (in Chinese)

Wu Z, Zhang Y, Chen Q, Wang H (2021). Attitude of Chinese public towards municipal solid waste sorting policy: A text mining study. Science of the Total Environment, 756: 142674

Yang H, Huang K, Zhang K, Weng Q, Zhang H, Wang F (2021a). Predicting heavy metal adsorption on soil with machine learning and mapping global distribution of soil adsorption capacities. Environmental Science & Technology, 55(20): 14316–14328

Yang J, Wen Y, Wang Y, Zhang S, Pinto J P, Pennington E A, Wang Z, Wu Y, Sander S P, Jiang J H, Hao J, Yung Y L, Seinfeld J H (2021b). From COVID-19 to future electrification: Assessing traffic impacts

on air quality by a machine-learning model. Proceedings of the National Academy of Sciences of the United States of America, 118(26): e2102705118

Zhong S, Zhang K, Bagheri M, Burken J G, Gu A, Li B, Ma X, Marrone B L, Ren Z J, Schrier J, Shi W, Tan H, Wang T, Wang X, Wong B M, Xiao X, Yu X, Zhu J J, Zhang H (2021). Machine learning: New ideas and tools in environmental science and engineering. Environmental Science & Technology, 55(19): 12741–12754

Zhu D, Zhou Q, Liu M, Bi J (2021). Non-optimum temperature-related mortality burden in China: Addressing the dual influences of climate change and urban heat islands. Science of the Total Environment, 782: 146760