Soil Environment Big Data: Development and Application Postprint

Authors: Guo Shuhai, Wu Bo, Zhang Lingyan, Luo Ming

Date: 2017-03-08T00:00:00+00:00

Abstract

Beginning with an analysis of big data characteristics, this article examines the development status of big data in the environmental field both domestically and internationally, clarifies the data foundation and bottleneck issues in the development of China' s soil environmental big data, and proposes construction methods and technical processes for soil environmental big data systems. Based on the national big data development strategy and industry demands in the soil environmental field, it recommends establishing in a coordinated manner soil environmental big data cloud platforms, management platforms, and thematic application platforms to provide public services and innovative application products oriented toward regional-scale soil environmental management, multi-agent cross-media collaborative governance, and agricultural product safety assurance.

Full Text

Preamble

Strategy & Policy Decision Research
ChinaXiv Partner Journal

Soil Environmental Big Data: Construction and Application

Guo Shuhai[1,2,3], Wu Bo[1,2,3], Zhang Lingyan[1,2,3], Luo Ming

Abstract

This paper begins with an analysis of big data characteristics and examines the development status of big data in the environmental field both domestically and internationally. It clarifies the data foundation and bottleneck issues in China' s soil environmental big data development, proposes construction methods and technical processes for soil environmental big data systems, and recommends establishing a unified soil environmental big data cloud platform, management

platform, and thematic application platform based on national big data development strategies and industry demands in the soil environment sector. These platforms would provide public services and innovative application products for regional-scale soil environmental management, multi-entity cross-media collaborative governance, and agricultural product safety assurance.

Keywords: soil environment, big data, digital management, multi-media synergetic remediation, agricultural product quality safety

DOI: 10.16418/j.issn.1000-3045.2017.02.011

1.1 Characteristics of Big Data

Big data refers to massive data collections whose scale in acquisition, storage, management, and analysis far exceeds the capacity of traditional database software tools. It features huge data volume, rapid data flow, and diverse data types, requiring new processing modes with stronger decision-making, insight discovery, and process optimization capabilities to be effectively utilized as an information asset [1,2]. Due to its enormous scale, big data exhibits two distinct characteristics compared to traditional data: (1) **Diverse data attributes**, including structured, semi-structured, and unstructured data [3,4]. Big data encompasses not only numbers but also text, images, audio, video, and other formats, covering rich content with strong mining potential and greater application value. (2) **Frequent data interaction**, with large-scale data analysis proceeding in parallel with real-time data mining [5]. In data analysis, structured data can follow established patterns [3], while analysis of semi-structured and unstructured data in big data follows unknown patterns that can only be explored through comprehensive simulation and hypothetical response approaches to calculate the credibility of various possibilities [3].

Big data acquisition primarily occurs in three forms: (1) collecting public information for trend judgment and customized services; (2) collecting sensor data for professional predictive analysis; and (3) collecting and integrating comprehensive data for correlation comparisons.

2 Development Status of Environmental Big Data

Environmental big data is currently in a vigorous development stage, demonstrating broad application prospects. Big data technology primarily encompasses data management, computational processing, and data analysis, with data analysis being the core. Data analysis has evolved through several historical stages: the first stage involved naive data analysis such as divination and agricultural calculations; the second stage featured mathematics-based data analysis using probability theory and statistics with computer technology; the third stage, following the information technology revolution, involved structured and digital data processing for integrated analysis based on computers and mathematics; and the fourth stage—current big data analysis—represents a

fusion of internet, automation, computer, and mathematical technologies. Thus, data analysis in big data technology is a broad concept that includes not only narrow-sense data analysis but also deep mining of massive datasets.

2.1 Rapid Development of Environmental Big Data in Europe and America

Due to higher informatization levels and better big data infrastructure, European and American countries have developed environmental big data more rapidly. The U.S. Environmental Protection Agency (EPA) has particularly applied environmental big data services to monitoring networks, data sharing, and public services [7,8]. In monitoring network construction, the EPA registers facilities with pollution discharge rights—including enterprises, wastewater treatment plants, civil facilities, and mining operations—building a discharge facility registration database through unique "facility identification codes" to enable cross-system and cross-database retrieval [8]. For data sharing, the EPA implements rapid, effective, secure, and accurate real-time environmental data exchange through the Central Data Exchange, connecting U.S. federal and local governments, enterprises, and EPA branches [9]. In public services, the EPA' s Envirofacts database [8] systematically opens air, water, waste, toxic substances, radiation, and soil environmental data to the public through map-based visualization, allowing searches for information on exhaust emissions, discharge permits, hazardous waste treatment processes, toxic chemical releases, and Superfund site status.

2.2 China' s Atmospheric Environment Management Leads in Big Data Adoption

Atmospheric environmental data is relatively easy to collect and analyze, and China' s urgent need for haze control has accelerated big data development in this sector. The Beijing Municipal Environmental Protection Bureau collaborated with IBM to develop an air quality prediction and modeling system based on cognitive computing, big data analytics, and IoT technology. By analyzing real-time data streams from air monitoring stations and meteorological satellites, and leveraging self-learning capabilities and supercomputing power, the system provides high-precision 72-hour air quality forecasts for the Beijing area, enabling real-time monitoring of pollution sources and distribution—known as the "Green Horizon" project [9]. This initiative uses big data and artificial intelligence to predict air pollution conditions up to 10 days in advance, allowing urban managers to take targeted measures such as adjusting traffic patterns and controlling industrial emissions in advance. With accurate predictions, the system can further collect unstructured data through apps, including weather patterns, scientific journal content, and government reports, evolving into cognitive technology.

Due to differences in environmental media, pollutant characteristics, monitoring methods, and historical accumulation, big data applications and prospects

vary across environmental sectors (Table 2). Atmospheric, water, and soil environmental big data each have distinct development characteristics, requiring targeted research on system construction and applications.

3 Development Status of Soil Environmental Big Data

3.1 Characteristics of Soil Environmental Big Data

Environmental research objects have different attributes, resulting in significant variations in available data types. Atmospheric environmental data can be collected at high frequencies through sensors, and the public has direct, sensitive perception of air quality. In contrast, soil environmental quality changes slowly with minimal fluctuations, and pollution exhibits cumulative and lagged characteristics. The public lacks direct sensory judgment capability, and automatic online monitoring is difficult, with higher costs for manual sampling and monitoring, making forecasting and early warning more challenging. However, these characteristics also provide an advantage for big data development: exploring causal relationships between soil environmental quality and various influencing factors based on the "source-sink" nature of soil environments. Through diversified data integration—such as spatial distribution data of pollution sources, pollutant emission categories and totals, multi-dimensional pollution pathways, environmental carrying capacity and spatial variations, and background value atlases and remote sensing images—multi-dimensional spatiotemporal big data models can be established.

3.2 Foundation for Soil Environmental Big Data Development

Since the 1980s, China has conducted multiple national-scale soil environmental surveys, including national background value investigations [10], soil pollution status surveys [11], multi-purpose geochemical surveys [12], and agricultural product origin environment surveys [13], generating over two million research papers and reports. These efforts have established soil environmental foundation databases and derivative databases focusing on agricultural land, contaminated sites, and drinking water source areas, involving local agricultural product and population health information (Table 3). In terms of data volume, these have essentially reached big data scales, though effective data extraction and deep mining are still required. The State Council' s *Soil Pollution Prevention and Control Action Plan* issued in 2016 [14] prioritizes soil pollution investigation and monitoring, establishing a system for regular soil environmental quality surveys every 10 years and building a soil environmental quality monitoring network to achieve full county-, city-, and district-level coverage by 2020. This provides nationwide foundational data sources for soil environmental big data, laying the groundwork for systems with massive sample sizes, multi-source data, and dynamic indicators. Building on this foundation, utilizing "Internet+" information exchange models for soil environmental data ingestion and supplementation through self-comparison, self-updating, and self-improvement will enable the construction of a distinctive Chinese soil environmental big data system.

This will achieve digital soil environment management and provide "targeted" decision-making solutions for regional and national-scale soil environmental protection and risk control.

3.3 Bottlenecks in Soil Environmental Big Data Development

Several problems hinder soil environmental big data development: (1) High costs and long cycles for soil environmental quality monitoring result in insufficient data accumulation; (2) China' s environmental monitoring system is still under construction with relatively single data types and elementary analysis methods, lacking data fusion and deep mining approaches and urgently needing mathematical models for correlation analysis; (3) Soil environmental quality management must be based on Geographic Information Systems (GIS), but limited extension capabilities of GIS tools and relational database management systems, constrained by data storage patterns, result in low efficiency in GIS spatial data automatic synthesis; and (4) The client-server architecture of GIS leads to weak capabilities in data sharing, storage, synchronous updating, and update efficiency. Therefore, technical integration should establish a data-driven, multi-industry, multi-disciplinary cross-fusion system for mutual benefit, forming an intelligent soil environmental management data support system.

4 Construction of Soil Environmental Big Data Systems

Big data features massive volume, diversity, and rapid change, while also having low value density, requiring data aggregation and extraction preprocessing for specific problems to reduce computational costs. Big data analysis project experience indicates that highly available, scalable data storage architectures and flexible, efficient data analysis frameworks are fundamental to building robust big data analysis systems. Due to diverse soil environmental big data acquisition channels and wide-ranging data sources, **data blending** must first be performed before integrated analysis. Data blending aims for intelligent decision-making by extracting, fusing, and integrating relevant data from multiple sources into an analytic dataset [15]. This analytic dataset is an independent and flexible entity that can be reorganized, adjusted, and updated as data sources change.

Multi-source data in the blending process [15] comes from three aspects: (1) **Primary data**, mainly internal data directly collected and controlled by project organizers; (2) **Secondary data**, mainly external data collected, organized, and provided by third parties; and (3) **Scientific data**, mainly obtained through scientific research, formula calculations, and model estimations. These three data types provide different information for system establishment. In big data analysis projects, data scientists need to collect, organize, and fuse relevant data from these three categories for specific problems.

Big data blending and system construction involve five basic steps: (1) extracting data from multiple heterogeneous sources; (2) organizing and classifying data; (3) cleaning data; (4) combining multi-source data, transforming it, and es-

tablishing datasets; and (5) building data analysis models for specific problems. According to the characteristics of soil environmental big data, the technical route for establishing a big data system centered on soil environmental quality should follow the process shown in Figure 1 [Figure 1: see original paper].

In this system, direct data refers to data that directly characterizes soil environmental quality, such as pollutant types, total amounts, and available content. Relevant data refers to data that influences soil environmental quality, including soil physicochemical properties, spatial distribution and emission characteristics of pollution sources, pollutant diffusion pathways, self-purification capacity of soil environments, and environmental quality characteristics of related media such as water and air, as well as meteorological data, hydrogeological data, environmental imagery, and remote sensing data in other formats.

5 Development Directions and Applications of Soil Environmental Big Data

Through digital soil environmental big data collections, thematic platforms for protection and prevention can be built to provide public services based on soil environmental big data. Using deep mining and knowledge discovery from big data, quantitative soil environmental management and multi-entity cross-media collaborative governance can be achieved. Targeted remediation and safe utilization of contaminated soil can establish digital traceability networks for agricultural product quality safety, thereby ensuring regional agricultural product quality and safety (Figure 2 [Figure 2: see original paper]).

5.1 Providing Digital Public Services Based on Soil Environmental Big Data

The State Council' s *Outline for Promoting Big Data Development* [16] calls for developing big data applications in industries such as manufacturing, emerging industries, and agriculture and rural areas to form a big data product system and improve the big data industry chain. The *Soil Pollution Prevention and Control Action Plan* [14] also requires using data from environmental protection, land resources, agriculture, and other departments to establish a soil environmental foundation database and build a national soil environmental information management platform. By leveraging mobile internet and IoT technologies, data acquisition channels should be broadened to enable dynamic updates. Accordingly, multi-source data fusion and digital characterization should be conducted to explore soil environmental quality databases and multi-assessment methods, regional analysis and target control models for soil environmental quality, and scenario analysis and decision-making technologies for contaminated soil remediation. A panoramic soil environmental quality analysis model should be established, and based on the needs of the soil environmental big data system, a unified soil environmental big data cloud platform and thematic application platform should be built to provide various digital public services based on soil

environmental data.

5.2 Conducting Multi-entity Cross-media Collaborative Governance at Regional Scales

Cross-media environmental pollution research represents the most active frontier field internationally, making it crucial to understand the sources, causes, impacts, and control of multi-media environmental pollution. Focusing solely on soil pollution prevention, risk control, and remediation is no longer sufficient to meet societal needs. There is an urgent need to strengthen correlation analysis of diversified data on pollution sources, pathways, and environmental carrying capacity for comprehensive assessment, forming national or cross-regional management platforms across ministries and industries. Therefore, based on traditional environmental management, data resources including economic and social development, basic geography, meteorology, and hydrology should be integrated to build a soil environmental big data system based on spatial geographic information systems to serve regional cross-media collaborative governance. For urban "brownfields" (abandoned land), cloud service platforms integrating basic data and information should be established for historical investigation of suspected contaminated sites, industrial analysis, multi-media interactions, and environmental countermeasures. For large-scale soil environmental management, information on water-soil-atmosphere environmental monitoring, mineral resource surveys, environmental capacity analysis, regional social development status, and industrial structure should be integrated to implement zoned, classified, and graded protection and governance [17].

5.3 Establishing Digital Traceability Networks for Agricultural Product Quality Safety

Soil environmental quality at agricultural product origins directly affects food safety. In China, soil pollution is severe in mid-south and southwest high-background-value regions, non-ferrous metal mining areas, large northern sewage irrigation districts, and urban suburbs of the Yangtze River Delta, Pearl River Delta, and Beijing-Tianjin-Hebei region, seriously threatening grain and vegetable quality safety. Therefore, establishing a precise plot-level agricultural product origin management platform through coding systems for risk early warning is an inevitable trend for future agricultural land soil environmental management. This platform would provide services for value-added sales of high-quality agricultural products and safety risk control of ordinary products. Current agricultural product traceability methods only allow post-event processing and will gradually be replaced by or integrated with pre-event intervention models. Thus, deep mining of existing soil environmental and agricultural product quality survey data to develop forecasting and decision-making technologies centered on collaborative control of heavy metal exceedance risks in agricultural products will become a major foundational task in the coming decade.

6 Recommendations

1. Promote comprehensive integration and sharing of soil environmental data resources, coordinate information technology project management, and establish free foundational state-managed information databases and fee-based commercial enterprise databases to eliminate data silos.
2. Establish an environmental information resource center to achieve data interconnection, forming a data sharing mechanism primarily based on direct platform acquisition and supplemented by inter-departmental data exchange.
3. Develop soil environmental big data analysis technologies to provide public services and commercial products, offering data and decision-making support for regional-scale soil environmental management, multi-entity cross-media collaborative governance, and agricultural product quality safety assurance.

References

1. McKinsey Global Institute. Big data: The next frontier for innovation, competition, and productivity. [2012-10-02]. http://www.mckinsey.com/Ingishts/MGI/Reaearch/Techno and-Innovation/Big-data-The-next-frontier-for-innovation
2. Viktor Mayer-Schönberger, Kenneth Cukier. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Translated by Sheng Yangyan, Zhou Tao. Hangzhou: Zhejiang People' s Publishing House, 2013.
3. Karger D R, Bakshi K, Huynh D, et al. Haystack: A customizable general-purpose information management tool for end users of semistructured data. *Proc. of the CIDR Conf*, 2005.
4. IBM. What is big data?. [2012-10-02]. http://www-01.ibm.com/software/data/bigdata/
5. Barwick H. The "four Vs" of Big Data. Implementing Information Infrastructure Symposium. [2012-10-02]. http://www.computerword.com.au/article/396198/iiis-four-vs-big-data/
6. Xu Guohu, Sun Ling. Online and offline e-commerce user data mining based on big data technology. *China Collective Economy*, 2012, (30): 187-188.
7. USEPA. Facility Registry Service (FRS). [2016-11-29]. https://www.epa.gov/enviro/facility-registry-service-frs
8. USEPA. Envirofacts. [2016-10-21]. https://www3.epa.gov/enviro/
9. Huanqiu Technology. IBM launches "Green Horizon" plan to help China' s "war on smog" . [2014-7-8]. http://tech.huanqiu.com/it/2014-07/5051822.html
10. Wei Fusheng, Yang Guozhi, Jiang Dezhen. Basic statistics and characteristics of soil element background values in China. *China Environmental Monitoring*, 1991, 7(1): 1-6.
11. Ministry of Environmental Protection, Ministry of Land and Resources. National Soil Pollution Survey Bulletin. [2014-4-17].

http://www.gov.cn/foot/2014-04/17/content_2661768.htm
12. Xi Xiaohuan. Multi-purpose regional geochemical survey. "*Tenth Five-Year*" *Important Geological Science and Technology Achievements and Major Prospecting Achievements Exchange Conference*, 2006.
13. Sina Finance. Ministry of Agriculture spends 5 years investigating soil conditions at agricultural product origins; these areas are severely polluted. [2016-11-25]. http://finance.sina.com.cn/roll/2016-11-25/docifxyawxa2737292.shtml
14. State Council. Soil Pollution Prevention and Control Action Plan. [2016-05-31]. http://www.gov.cn/zhengce/content/2016-05/31/content_5078377.htm
15. Wang Shan, Wang Huiju, Qin Xiaolin, et al. Architecture and implementation of big data analysis systems. *Big Data*, 2014, 2(1): 46-56.
16. State Council. Outline for Promoting Big Data Development. [2015-9-5]. http://www.gov.cn/zhengce/content/2015-09/05/content_10137.htm
17. Guo Shuhai, Wu Bo, Zhang Lingyan, et al. *Contaminated Site Risk Management and Remediation*. Beijing: Science Press, 2014.

Guo Shuhai is a professor and principal research scientist at the Institute of Applied Ecology, Chinese Academy of Sciences, and Director of the National-Local Joint Engineering Laboratory of Contaminated Soil Remediation by Biophysicochemical Synergistic Process. He is a core member of the CAS Distinguished Researcher Program. His research focuses on risk assessment and remediation of contaminated soils, and soil environmental data blending and mining. In recent years, he has led 11 projects including the "973" and "863" programs, national science and technology major projects, international cooperation projects, and national public welfare projects, receiving 9 provincial or ministerial science and technology awards.
E-mail: shuhaiguo@iae.ac.cn

Note: Figure translations are in progress. See original paper for figures.