China industrial environmental database 1998–2015

Haoqi Qian1,2,10✉, Feizhou Ren3,4,10, Yanran Gong5, Rong Ma4, Wendong Wei6,7 & Libo Wu4,8,9✉

There has been a rapid-growing trend in studying China’s environmental problems in the past decade. However, the existing environmental statistics data are far from meeting researchers’ requirements. The biggest problem is that the official environmental statistics data are only provided at either regional level or sectoral level. Considering the huge heterogeneities in different regions and sectors, researchers are unable to conduct comprehensive policy evaluations. In this study, we constructed the time-series industrial environmental database for China (CIED) at both regional and sectoral level. The database includes totally 31 regions and four types of pollutants: chemical oxygen demand (COD), sulphur dioxide (SO2), ammonia-nitrogen (NH3-N), and nitrogen oxide (NO\textsubscript{X}). This study also clarifies several important concepts for researchers to better understand China’s official environmental statistics data.

Background & Summary

China’s environmental problem is one of the most important issues accompanying with its rapid industrialization. Although the situation has been greatly improved in the past decade, the whole nation, especially in certain regions, is still facing severe environmental problems such as air pollution issue, water pollution and so on. Since 2000s, there is a growing number of evidence-based literature focusing on China’s environmental problems. Topics of these literature include discussions of China’s environmental Kuznets curve1–5, analyses of environmental protection performance6–8 and evaluations of environmental policies9,10. Most of these studies use either regional-level or sector-level environmental data that are retrieved from official statistics to conduct analyses, which fail to take both regional and sectoral heterogeneities into consideration simultaneously. The main reason is that the official statistics such as China Statistical Yearbook on Environment only report provincial aggregated pollution data and sectoral aggregated pollution data. Attempts to partly overcome this problem are to use micro-level data such as firm-level or plant-level data, and cases can be seen in several recent studies12–15. These studies provide insights into firms’ polluting behaviours and also do provide new evidence in understanding regional and sectoral variations in China.

There are two main drawbacks in using micro-level data, especially firm-level data, to study the environmental issues. Firstly, a certain number of observations will be dropped when constructing the balanced panel dataset which plays an important role in most empirical studies. Sample representativeness will be further weakened if researchers match different firm-level datasets in order to obtain more variables. As a result, other important topics such as structure effect, entry effect and exit effect can not be investigated at the macro-level. Secondly, many important influence factors are only available and applicable in regional level or sectoral level, and micro-level data only contains individual firm specific characteristics. But most of current research interests and policy needs are at the macro-level. In this case, firm-level analysis may unable to deliver desired results and policy implications at the macro-level. Due to the above two reasons, estimating environmental data at both regional level and sectoral level will undoubtedly contribute to the existing empirical works. For example, the two-dimensional data have great potentials to boost environmental policy evaluations by using treatment effect

1Institute for Global Public Policy and MOE Laboratory for National Development and Intelligent Governance, Fudan University, Shanghai, 200433, China. 2LSE-Fudan Research Centre for Global Public Policy, Fudan University, Shanghai, 200433, China. 3School of Economics, Fudan University, Shanghai, 200433, China. 4School of Data Science, Fudan University, Shanghai, 200433, China. 5School of International and Public Affairs, Shanghai Jiao Tong University, Shanghai, 200030, China. 6SJTU-UNIDO Joint Institute of Inclusive and Sustainable Industrial Development, Shanghai Jiao Tong University, Shanghai, 200030, China. 7Institute for Big Data, Fudan University, Shanghai, 200433, China. 8Shanghai Institute for Energy and Carbon Neutrality Strategy, Fudan University, Shanghai, 200433, China. 9Institute for Global Public Policy and MOE Laboratory for National Development and Intelligent Governance, Fudan University, Shanghai, 200433, China. 10These authors contributed equally: Haoqi Qian, Feizhou Ren. ✉e-mail: qianhaoqi@fudan.edu.cn; wulibo@fudan.edu.cn
analyses such as difference-in-difference (DID) model\textsuperscript{16–18} and synthetic control method (SCM)\textsuperscript{19,20}, which have gained much popularity in this field in the past decade. The second direction of the application is conducting environmental efficiency analyses such as data envelope analysis (DEA) and stochastic frontier analysis (SFA), by viewing pollutants as undesirable products\textsuperscript{21–24}. Furthermore, two-dimensional environmental data can also be used to construct environmentally extended multi-regional input-output (EE-MRIO) tables\textsuperscript{25–27}, as well as provide more accurate estimates for parameters\textsuperscript{28}, which are extremely useful in conducting complex economic system analyses such as computable general equilibrium (CGE) modelling works.

In the China Industrial Environmental Database (CIED), there are 31 regions (excluding Hongkong, Macao, and Taiwan) and 39 industrial sectors. For each sector in each region, we provide environmental data of four types of pollutants: chemical oxygen demand (COD), sulfur dioxide (SO\textsubscript{2}), ammonia-nitrogen (NH\textsubscript{3}-N), and nitrogen oxide (NO\textsubscript{X}). The database contains discharge and removal data for all four pollutants. Due to the data availability, data for COD and SO\textsubscript{2} are available from 1998 to 2015, data for NH\textsubscript{3}-N is available from 2001 to 2015 and data for NO\textsubscript{X} is available from 2006 to 2015. The time period covers last three years of the ninth Five-Year Plan and the successive three complete Five-Year Plan periods. Since Five-Year Plans play great roles in the policy-making procedure, the database can be used to analyze a wide range of institutional and policy adjustment issues during this period.

**Methods**

In this study, we combine the bottom-up method and top-down method to construct the two-dimensional environmental database. In the first step, the bottom-up method uses firm-level environmental data calculate pollutants’ discharge and removal amount by region and by sector. In the second step, the top-down method collects total amount of pollutants discharged and removed for each region or for each sector. In the final step, the cross-entropy method is used to balance the two-dimensional matrix in each year.

**Aggregation of micro-level environmental data.** Micro-level environmental data are retrieved from China’s Environmental Statistics Database (CESD). The CESD is a micro-level database compiled by China’s Ministry of Environmental Protection (MEP) which is used as the data basis for publications of China’s official environmental statistics such as China Statistical Yearbook on Environment, China Environment Yearbook and Annual Statistic Report on Environment in China. This study uses the industrial firm-level data from the CESD and we call it China’s Industrial Environmental Statistics Database (CIESD). The CIESD is a newly released database available to researchers which covers all major industrial emission sources in China. This database has already been used and explained in details in several recent studies\textsuperscript{6,15,28,30}.

Table 1 reports the number of firms covered in our database by region and year. We should pay some special attentions when using the database. Totally 20 cells in Table 1 are zero or close to zero for all regions, which means these observations are completely or almost missing. The overall data coverage rates in 1998 and 1999 are 75.38% and 91.97% respectively, which are not so perfect but acceptable for this study. In 2006, there are totally 1784 firms which cannot be categorized to any region by using 6-digit administrative division code. After carefully checking these observations manually, we find they are all duplicated observations for Anhui province. After comparing data of these duplicated observations, we just keep one unique observation with useful information. For the period during 2006 and 2010, environmental statistical data for thermal power plants are collected as a separate database, which is not included in our database. Total number of thermal power plants are also reported in Table 1. Therefore, if we subtract numbers of thermal power plants in the period during 2006 and 2010, the overall data coverage rates from 2000 to 2014 are quite close to 100% which means a quite good sample representativeness.

Before aggregating the micro-level environmental data, deeper investigations of the whole dataset have been conducted in order to revise the abnormal data records and improve the data quality. There are typically two types of errors exist for the abnormal data record. The first type is missing data error and the other one is measurement error. The missing data error will lead to underestimation of pollutants’ values and the measurement error will lead to both underestimation and overestimation of pollutants’ values. We manually checked each firm’s time series data to revise these two issues. For the missing data, if a water-polluting firm has both COD and NH\textsubscript{3}-N discharge data in all year but lacks NH\textsubscript{3}-N discharge data in one specific year, then the missing NH\textsubscript{3}-N discharge data is identified and will be filled with the interpolation value. For the measurement error, it is likely that firms may report incorrect data by using different units. For example, if one firm reports the data in unit of gram instead of kilogram which is required by the regulations, then data value recorded in the database will be 1,000 times higher than the true value. This data will be identified as abnormal data and be revised to its reasonable level.

The raw two-dimensional environmental data can be calculated based on revised CIESD as follows:

$$X_{rj}^c = \frac{\sum_i X_{rj,ci}^c}{X_{rj}^c}$$

(1)

where $X_{rj,ci}^c$ is revised environmental data of firm $i$ of sector $j$ in region $r$. $X_{rj}^c$ represents raw macro-level environmental data of sector $j$ in region $r$. Superscript $c$ represents type of indicator, i.e. discharge or removal. We estimate the missing $X_{rj}^c$ by extrapolating the ratio of each sector among all regions, and the ratio is defined as follows:

$$\text{Ratio}_{rj}^c = \frac{X_{rj}^c}{\sum_r X_{rj}^c}$$

(2)
### Table 1. Distribution of CIESD firms across 31 regions, 1998–2014. Data for CIESD are collected from China Statistical Yearbook on Environment and China Environment Yearbook. Total firm numbers are retrieved from tables grouped by sectors. Total number of CIESD thermal power plants are collected from Annual Statistical Report on Environment in China.

Since micro-level environmental data in 2015 is not available, we use the structure ratios of 2014 as the prior structure information for 2015.

**Collection of macro-level environmental data.** In this study, we collect macro-level environmental data from China Statistical Yearbook on Environment for year from 2001 to 2015 and China Environment Yearbook for year from 1998–2000. Environmental data in the yearbook are compiled from the CIESD, which is collected and processed according to the environmental statistical system (ESS). The ESS was adjusted at the beginning of each Five-Year Plan period to improve the support the environmental statistical work and to improve the quality of environmental statistical data. Within the time period of our database, the ESS has been adjusted four times. These adjustments raise three biggest challenges for this study to construct the CIESD.

The first challenge is that not all pollutants are included in the CIESD from 1998 to 2015. For example, NH$_3$-N was reported since the Tenth Five-Year Plan (started from 2001) and NO$_x$ was reported since the Eleventh Five-Year Plan (started from 2006). Due to the data availability, environmental data of these two pollutants are...
| Data type | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|-----------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| COD       |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| Discharge | 1.00 | 1.01 | 1.01 | 1.15 | 1.13 | 1.16 | 1.13 | 1.12 | 1.17 | 1.13 | 1.16 | 1.19 | 1.10 | 1.11 | 1.12 | 1.13 | 1.15 |
| Removal   | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| Production|      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| SO₂       |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| Discharge | 1.00 | 1.02 | 1.01 | 1.11 | 1.11 | 1.20 | 1.08 | 1.10 | 1.09 | 1.09 | 1.08 | 1.10 | 1.09 | 1.07 | 1.09 | 1.10 | 1.11 |
| Removal   | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.01 | 1.00 | 1.00 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| Production|      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| NH₃-N     |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| Discharge | 1.14 | 1.11 | 1.09 | 1.09 | 1.10 | 1.13 | 1.11 | 1.11 | 1.12 | 1.11 | 1.11 | 1.07 | 1.09 | 1.10 | 1.11 | 1.11 |      |
| Removal   | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |      |
| Production|      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| NOₓ       |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| Discharge | 1.09 | 1.09 | 1.09 | 1.07 | 1.07 | 1.07 | 1.07 | 1.07 | 1.07 | 1.07 | 1.07 | 1.04 | 1.05 | 1.06 | 1.06 | 1.06 | 1.08 |
| Removal   | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |      |
| Production|      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |

Table 2. Brief summary of pollutants during four Five-Year Plan periods. Information of key pollutants are collected by authors from China's Total Emission Control of Major Pollutants in each Five-Year Plan period.

The third challenge is that the classifications of sectors reported in official statistics have been changed three times within the data period. Classifications of sectors are based on Classification Standards of National Economic Industries (CSNEI) which includes 2-digit, 3-digit and 4-digit codes for all sectors and sub-sectors. Three versions of CSNEI used for classification are GB/T 4754-1994, GB/T 4754-2002 and GB/4754-2011. In the period from 1998 to 2000, all firms are classified into 20 sectors (including cement manufacturing as a sub-sector). The classification changed for the first time in 2001 and 2002, all firms are classified into 43 sectors (including cement manufacturing and thermal power as sub-sectors). Then the classification changed for the second time from 2003 to 2010, all firms are classified into 42 sectors (including cement manufacturing and thermal power as sub-sectors). Finally, the classification changed for the third time from 2011 to 2015, all firms are classified into 42 sectors (including four separate sub-sectors with regional data). To make the data comparable across different years, we coordinate all sectors into 39 sectors labelled from 1 to 39. Table 4 reports the concordance of industrial sectors and the first column is the coordinated sector number. We also find that for some firms, they have different 2-digit sector codes recorded in CIESD and Annual Survey of Industrial Enterprises Database (ASIED). Since 2-digit sector codes in ASIED are more consistent to the sector classifications of official statistics released by the National Bureau of Statistics of China. We match the firms in CIESD and ASIED, and revise firms’ 2-digit sector codes according to the ASIED before the concordance of sectors. Besides, firms classified as “Others” in CIESD are all checked and revised manually. The chord diagram in Fig. 1(a) illustrates the concurrence of sectors for the whole period from 1998 to 2015.

The second challenge is that indicators of pollutants are not consistent within the whole data period. According to the ESS, industrial emission sources are divided into major sources and non-major sources. Only major sources are covered by the CESD, and they are expected to account for at least 85% of total annual discharges of major pollutants. When compiling the environmental statistical data from 1998 to 2010, regional removal data, sectoral discharge data and sectoral removal data are just the summation value of major sources. However, regional discharge data are adjusted to include non-major sources’ discharge data which are estimated by using certain estimation methods. As a result, the national data aggregated from regional data are typically larger than that aggregated from sectoral data. Besides, removal data are no longer reported according to the new ESS since the beginning of Twelfth Five-Year Plan. Instead, number of pollutants produced are reported in the CESD as well as in the yearbooks. Both the new production data and the discharge data are adjusted to include non-major sources. Table 3 reports the comparison results of regional aggregation to sectoral aggregation for three indicators of four pollutants, 1998–2015. Raw data are collected from China Statistical Yearbook on Environment and China Environment Yearbook. Ratios are calculated by authors.
| No | Concordant sector | 1998–2000 Sectors | 2-digit code (1994) | 2001–2002 Sectors | 2-digit code (1994) | 2003–2010 Sectors | 2-digit code (2002) | 2011–2014 Sectors | 2-digit code (2011) |
|----|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
| 1  | Mining and Washing of Coal | Mining and quarrying (part) | 06 | Coal Mining and Dressing | 06 | Mining and Washing of Coal | 06 |
| 2  | Extraction of Petroleum and Natural Gas | | 07 | Petroleum and Natural Gas Extraction | 07 | Extraction of Petroleum and Natural Gas | 07 |
| 3  | Mining and Processing of Ferrous Metal Ores | | 08 | Ferrous Metals Mining and Dressing | 08 | Mining and Processing of Ferrous Metal Ores | 08 |
| 4  | Mining and Processing of Non-ferrous Metal Ores | | 09 | Nonferrous Metals Mining and Dressing | 09 | Mining and Processing of Non-ferrous Metal Ores | 09 |
| 5  | Mining and Processing of Non-metal Ores | | 10 | Nonmetal Minerals Mining and Dressing | 10 | Mining and Processing of Non-metal Ores | 10 |
| 6  | Mining of Other Ores | | 11 | Other Minerals Mining and Dressing | 11 | Mining of Other Ores | 11 |
| 7  | Food Processing | Food, beverage, and tobacco processing | 13 | Food Processing | 13 | Processing of Food from Agricultural Products | 13 |
| 8  | Manufacture of Food Products | | 14 | Food Manufacturing | 14 | Manufacture of Foods | 14 |
| 9  | Manufacture of Wine, Drinks and Refined Tea | | 15 | Beverage Manufacturing | 15 | Manufacture of Beverages | 15 |
| 10 | Manufacture of Tobacco | | 16 | Tobacco Processing | 16 | Manufacture of Tobacco | 16 |
| 11 | Manufacture of Textile | Textile Industry | 17 | Manufacture of Textile | 17 | Manufacture of Textile | 17 |
| 12 | Manufacture of Textile Wearing and Apparel | | 18 | Garments and Other Fiber Products | 18 | Manufacture of Textile Wearing Apparel, Footwear, and Caps | 18 |
| 13 | Manufacture of Leather, Fur, Feather and Related Products and Footwear | | 19 | Leather, Furs, Down and Related Products | 19 | Manufacture of Leather, Fur, Feather and Related Products | 19 |
| 14 | Processing of Timber, Manufacture of Wood, Bamboo, Rattan, Palm, and Straw Products | Other Industries (part) | 20 | Timber Processing, Bamboo, Cane, Palm Fiber and Straw Products | 20 | Processing of Timber, Manufacture of Wood, Bamboo, Rattan, Palm, and Straw Products | 20 |
| 15 | Manufacture of Furniture | | 21 | Furniture Manufacturing | 21 | Manufacture of Furniture | 21 |
| 16 | Manufacture of Paper Products | Papermaking and Paper Products | 22 | Papermaking and Paper Products | 22 | Manufacture of Paper and Paper Products | 22 |
| 17 | Printing, Reproduction of Recording Media | | 23 | Printing | 23 | Printing, Reproduction of Recording Media | 23 |
| 18 | Manufacture of Articles for Culture, Education and Sport Activity | Other Industries (part) | 24 | Cultural, Educational and Sports Goods | 24 | Manufacture of Articles for Culture, Education and Sport Activity | 24 |
| 19 | Processing of Petroleum, Coking, Processing of Nuclear Fuel | Petroleum Processing and Coking | 25 | Petroleum Processing and Coking | 25 | Processing of Petroleum, Coking, Processing of Nuclear Fuel | 25 |
| 20 | Manufacture of Raw Chemical Materials and Chemical Products | Raw Chemical Materials and Chemical Products | 26 | Raw Chemical Materials and Chemical Products | 26 | Manufacture of Raw Chemical Materials and Chemical Products | 26 |
| 21 | Manufacture of Medicines | Medical and Pharmaceutical Products | 27 | Medical and Pharmaceutical Products | 27 | Manufacture of Medicines | 27 |
| 22 | Manufacture of Chemical Fibers | Chemical Fiber | 28 | Chemical Fiber | 28 | Manufacture of Chemical Fibers | 28 |
| 23 | Manufacture of Rubber and Plastic | Rubber Products | 29 | Rubber Products | 29 | Manufacture of Rubber | 29 |
|   |                       | Plastic Products | 30 | Plastic Products | 30 | Manufacture of Plastic | 30 |

Continued
| Concordance sector | 1998–2000 Sectors | 2-digit code (1994) | 2001–2002 Sectors | 2-digit code (1994) | 2003–2010 Sectors | 2-digit code (2002) | 2011–2014 Sectors | 2-digit code (2011) |
|--------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
| 24                 | Manufacture of Non-metallic Mineral Products | 31 | Nonmetal Mineral Products | 31 | Manufacture of Non-metallic Mineral Products | 31 | Manufacture of Non-metallic Mineral Products | 30 |
| 25                 | Smelting and Pressing of Ferrous Metals | 32 | Smelting and Pressing Of Ferrous Metals | 32 | Smelting and Pressing of Ferrous Metals | 32 | Smelting and Pressing of Ferrous Metals | 31 |
| 26                 | Smelting and Pressing of Non-ferrous Metals | 33 | Smelting and Pressing Of Non-ferrous Metals | 33 | Smelting and Pressing of Non-ferrous Metals | 33 | Smelting and Pressing of Non-ferrous Metals | 32 |
| 27                 | Manufacture of Metal Products | 34 | Metal Products | 34 | Manufacture of Metal Products | 34 | Manufacture of Metal Products | 33 |
| 28                 | Manufacture of General Purpose Machinery | 35 | Ordinary Machinery Manufacturing | 35 | Manufacture of General Purpose Machinery | 35 | Manufacture of General Purpose Machinery | 34 |
| 29                 | Manufacture of Special Purpose Machinery | 36 | Special Purpose Equipment Manufacturing | 36 | Manufacture of Special Purpose Machinery | 36 | Manufacture of Special Purpose Machinery | 35 |
| 30                 | Manufacture of Automobile | 37 | Transport Equipment Manufacturing | 37 | Manufacture of Automobile | 37 | Manufacture of Automobile | 36 |
| 31                 | Manufacture of Railway, Shipbuilding, Aerospace and Other Transportation Equipment | 38 | Machine, Electric, Machinery and Electronic Equipment Manufacturing | 38 | Manufacture of Railway, Shipbuilding, Aerospace and Other Transportation Equipment | 38 | Manufacture of Railway, Shipbuilding, Aerospace and Other Transportation Equipment | 37 |
| 32                 | Manufacture of Electrical Machinery and Equipment | 40 | Electric Equipment and Machinery | 40 | Manufacture of Electrical Machinery and Equipment | 40 | Manufacture of Electrical Machinery and Equipment | 39 |
| 33                 | Manufacture of Computers, Communication, and Other Electronic Equipment | 41 | Electronic and Telecommunications Equipments | 41 | Manufacture of Communication Equipment, Computers and Other Electronic Equipment | 41 | Manufacture of Communication Equipment, Computers and Other Electronic Equipment | 39 |
| 34                 | Manufacture of Measuring Instrument | 42 | Instrument, Meters, Cultural and Office Machinery Manufacturing | 42 | Manufacture of Measuring Instruments and Machinery for Cultural Activity and Office Work | 42 | Manufacture of Measuring Instruments and Machinery for Cultural Activity and Office Work | 41 |
| 35                 | Production and Supply of Electric Power and Heat Power | 44 | Electric Power, Steam and Hot Water Productions | 44 | Production and Distribution of Electric Power and Heat Power | 44 | Production and Distribution of Electric Power and Heat Power | 44 |
| 36                 | Production and Supply of Gas | 45 | Gas Production and Supply | 45 | Production and Distribution of Gas | 45 | Production and Supply of Gas | 45 |
| 37                 | Production and Supply of Water | 46 | Tap Water Production and Supply | 46 | Production and Distribution of Water | 46 | Production and Supply of Water | 46 |
| 38                 | Other Manufactures | 43 | Other Manufactures | 43 | Manufacture of Artwork and Other Manufacturing | 42 | Other Manufactures | 41 |
| 39                 | Others | 99 | Others | 99 | Others | 99 | Others | 99 |

Table 4. Concordance of industrial sectors. The mapping rules are set by authors after comparing sector names and descriptions manually. Complete list of 4-digit sector code mapping rules used for concordance are provided in the Supplementary Table 1.

The changes of all firms’ sector codes in CIESD. We can see that most of firms’ sector codes in the CIESD are not revised or remain the same 2-digit sector codes after revision. If we take a close look at firms whose 2-digit sector codes are revised, interchanges occur in most sectors except for the “Others” sector whose 2-digit sector code is 39. Since all firms originally classified in the “Others” sector have been manually checked and revised, they are now assigned the correct 2-digit sector codes.
Moreover, since China has conducted the first National Census of Pollution Sources (NCPS) in 2007, the ESS has significantly changed according to the census results. As a result, there are two special concerns we must pay attention to when using the environmental data. Firstly, the ESS relied heavily on firms' self-reported data to determine the major emission sources before 2011. After adopting the complete emission source list provided by the NCPS, total number of firms included in the CESD increased significantly in 2011 and keeps stable thereafter. As a result, it is not surprising that total discharge data jump at 2011 as well. Secondly, a large number of discharge coefficients have been updated by the NCPS and they have been adopted in the ESS adjusted in 2011. Consequently, some sectors pollution levels also changed greatly. If researchers conduct empirical studies by using panel models, these systematic changes can be captured by the sector and time fixed effects. While if researchers use methods such as index decomposition, data envelope analysis and so on, the forementioned concerns should be treated carefully.

Balancing the environmental data. We use a two-step adjustment procedure to adjust the raw two-dimensional environmental data to construct the final CIESD. In the first step, the raw value is adjusted by using the following formula:

$$\tilde{X}_{r,j}^c = X_{r,j}^c \times \frac{Y_r^s}{\sum_j Y_{r,j}^s}$$

where $\tilde{X}_{r,j}^c$ represents the adjusted raw value and $Y_r^s$ represents aggregated macro-level environmental data for region $r$. The second term in the right-hand side of formula (3) is the region-specific average scale ratio to add non-major emission sources. Sectors in the same region are assumed to be scaled up proportionally. This step adjusts the aggregated regional raw data and make them consistent to the macro-level environmental data.

In the second step, we use the cross-entropy method to balance the two-dimensional data matrix. In this paper, we assume the regional distributions of the environmental data within each sector are more reliable. Thus, we define the prior coefficients $\tilde{a}_{r,j}^c$ and the new coefficients to be estimated $\tilde{a}_{r,j}^c$ as follows:

Table 5. The structure of the environmental data by region and by sector. The names of region 1 to 31, sector 1 to 39 can be found in Tables 1 and 4, respectively.

| Year | Sector 1 | Sector 2 | … | Sector 39 | Total |
|------|----------|----------|---|-----------|-------|
| Region 1 |          |          |   |           |       |
| Region 2 |          |          |   |           |       |
| … |          |          |   |           |       |
| Region 31 |          |          |   |           |       |
| Total |          |          |   |           |       |

Fig. 1 Revisions of firms’ 2-digit sector codes. (a) Whole dataset; (b) Only firms whose 2-digit sector codes are revised. Each node in the chord diagram represents one 2-digit sector and arcs represent directions and numbers of firms that have been revised from one sector to another.
| Data type | 1998  | 1999  | 2000  | 2001  | 2002  | 2003  | 2004  | 2005  | 2006  | 2007  | 2008  | 2009  | 2010  | 2011  | 2012  | 2013  | 2014  | 2015  |
|-----------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| COD       | 1.5060| 1.2658| 1.1984| 1.1092| 1.1520| 1.1472| 1.1301| 1.1161| 1.1172| 1.1201| 1.1202| 1.1231| 1.1201| 1.1538| 1.1576| 1.1419| 1.1669| 1.1743|
| Discharge | 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000|
| Removal   | 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000|
| SO2       | 1.4534| 1.9863| 1.9884| 1.9895| 1.9906| 1.9901| 1.9896| 1.9894| 1.9888| 1.9893| 1.9892| 1.9892| 1.9894| 1.9896| 1.9898| 1.9899| 1.9899| 1.9899|
| Discharge | 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000|
| Removal   | 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000|
| NH3-N     | 1.1200| 1.1120| 1.0792| 1.0841| 1.1057| 1.1213| 1.0861| 1.0792| 1.0847| 1.1048| 1.0643| 1.0815| 1.0842| 1.0874| 1.0903| 0.9996| 1.0000| 1.0025|
| Discharge | 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000|
| Removal   | 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000|
| NOX       | 1.0827| 1.0371| 1.0742| 1.0631| 1.0572| 1.0405| 1.0441| 1.0529| 1.0637| 1.0803| 1.0554| 1.0452| 1.0423| 1.0509| 1.0518| 1.0500| 1.0500| 1.0500|
| Discharge | 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000|
| Removal   | 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000| 0.0000|

Table 6. Estimated slope coefficients for different pollutants by year. Standard deviations are provided in the parenthesis.
where $\tilde{X}_{rj}^c$ is the new environmental data to be estimated. Then the objective function is defined as follows:

$$\min \sum_r \sum_j \alpha_{rj} \ln \frac{\tilde{X}_{rj}}{\bar{X}_{rj}}$$

subject to:

$$\sum_r \sum_j \left( \tilde{X}_{rj} - \bar{X}_{rj} \right) = \sum_r Y_r$$

$$\sum_r \tilde{X}_{rj} = 1 \text{ and } 0 \leq \tilde{X}_{rj} \leq 1$$

The solution for $\tilde{X}_{rj}$ is obtained by solving the optimization problem from formula (6)-(8) and the final estimated environmental data are obtained as follows:

$$\tilde{X}_{rj} = \tilde{X}_{rj} - \sum_r \bar{X}_{rj}$$

### Data Records

Our data records are available through Figshare in format of Excel file from the repository: [https://doi.org/10.6084/m9.figshare.16846966](https://doi.org/10.6084/m9.figshare.16846966). Table 5 presents the structure of the environmental data for each year by region and by sector. Each matrix includes 31 regions and 39 sectors. Totally 122 matrices are included in the database. Of these,

- 72 matrices are discharge and removal data for COD and SO$_2$ from 1998 to 2015;
- 30 matrices are discharge and removal data for NH$_3$-N from 2001 to 2015;
- 20 matrices are discharge and removal data for NO$_x$ from 2006 to 2015.

Unit for all environmental data is kilogram (kg).

### Technical Validation

**Creditability of firm-level data.** The prior information to construct the environmental data matrix are aggregated from the firm-level data, thus it is important to validate the creditability of these micro-level data. Since the yearbooks are the only authoritative data sources and sectoral data in the yearbooks only cover major emission sources, we compare data in our database to yearbooks’ data. In the Supplementary Fig. 1, we provided the comparison information for all pollutants by sector and by year. For years from 2000 to 2014, aggregation value for each sector in our database are very close to that in the yearbook. For years 1998 and 1999, although total observations are less than the yearbook, ratios that aggregation value in our database divided by total value in the yearbook are similar across sectors. Therefore, firm-level data in our database is credible and it is appropriate to calculate the coefficients in formula (4) by sector.

### Table 7. Adjusted overall R$^2$ for different pollutants by year.

| Data type | 1998      | 1999      | 2000      | 2001      | 2002      | 2003      | 2004      | 2005      | 2006      | 2007      | 2008      | 2009      | 2010      | 2011      | 2012      | 2013      | 2014      | 2015      |
|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| COD       | 0.9531    | 0.9883    | 0.9879    | 0.9970    | 0.9962    | 0.9960    | 0.9959    | 0.9976    | 0.9954    | 0.9983    | 0.9984    | 0.9973    | 0.9944    | 0.9991    | 0.9988    | 0.9986    | 0.9982    | 0.9973    |
|           | 0.9460    | 0.9986    | 0.9999    | 0.9990    | 0.9991    | 0.9888    | 0.9992    | 1.0000    | 0.9955    | 0.9999    | 1.0000    | 1.0000    | 1.0000    | 1.0000    | 1.0000    | 0.9996    | 0.9996    | 0.9982    |
| SO$_2$    | 0.9852    | 0.9947    | 0.9959    | 0.9992    | 0.9990    | 0.9972    | 0.9993    | 0.9977    | 0.9986    | 0.9997    | 0.9998    | 0.9996    | 0.9958    | 0.9995    | 0.9997    | 0.9996    | 0.9994    | 0.9991    |
|           | 0.9778    | 0.9938    | 0.9999    | 0.9986    | 0.9974    | 0.9975    | 0.9991    | 0.9933    | 0.9999    | 1.0000    | 1.0000    | 1.0000    | 1.0000    | 1.0000    | 1.0000    | 0.9992    | 0.9886    | 0.9991    |
| NH$_3$-N  | 0.9976    | 0.9883    | 0.9991    | 0.9980    | 0.9976    | 0.9886    | 0.9898    | 0.9889    | 0.9887    | 0.9974    | 0.9996    | 0.9993    | 0.9992    | 0.9991    | 0.9984    | 0.9992    | 0.9991    | 0.9984    |
|           | 0.9993    | 1.0000    | 1.0000    | 1.0000    | 0.9966    | 0.9995    | 0.9996    | 1.0000    | 1.0000    | 1.0000    | 0.9994    | 0.9998    | 0.9998    | 0.9998    | 0.9998    | 0.9998    | 0.9998    | 0.9998    |
| NO$_x$    | 0.9988    | 0.9961    | 0.9986    | 0.9998    | 0.9999    | 0.9999    | 0.9999    | 0.9999    | 0.9999    | 0.9999    | 0.9999    | 0.9999    | 0.9999    | 0.9999    | 0.9999    | 0.9999    | 0.9999    | 0.9994    |
|           | 0.9993    | 1.0000    | 1.0000    | 0.9981    | 1.0000    | 1.0000    | 0.9997    | 0.9999    | 0.9999    | 0.9999    | 0.9999    | 0.9999    | 0.9999    | 0.9999    | 0.9999    | 0.9999    | 0.9999    | 0.9996    |
Validation of balanced results. Our estimation procedure aims to adjust the environmental data and make them consistent to the regional aggregations. Since regional aggregations include non-major emission sources, estimated environmental data have been systematically scaled up. For each year, we run the following two-way fixed effect regression to investigate adjustments from the original data to the estimated data:

\[ \hat{X}_{rj} = \alpha_r + \gamma_j + \beta X_{rj} + \epsilon_{rj} \]  

Fig. 2 Comparison of unbalanced value and balanced value from 1998 to 2015. (a) COD discharge; (b) COD removal; (c) SO\(_2\) discharge; (d) SO\(_2\) removal; (e) NH\(_3\)-N discharge; (f) NH\(_3\)-N removal; (g) NO\(_x\) discharge; (h) NO\(_x\) removal. Fitted values are obtained from pooled OLS regressions.
Fig. 3  Comparison of Zhejiang’s official value and our value. (a) COD discharge; (b) SO₂ discharge; (c) NH₃-N discharge; (d) NOₓ discharge. Fitted values are obtained from pooled OLS regressions.

Fig. 4  Comparison of original and revised national aggregation data from 1998 to 2015. (a) COD; (b) SO₂; (c) NH₃-N; (d) NOₓ. Discrepancies are shown in bars on the secondary axis.
where \( \beta \) is the slope coefficient and \( \varepsilon_{rj}^c \) is the noise term which follows normal distribution, \( \varepsilon_{rj}^c \sim N(0, \sigma^2_c) \). Table 6 presents estimated slope coefficients and standard deviations for different pollutants. All coefficients are statistically significant at 1% level. Table 7 presents adjusted \( R^2 \) for different pollutants and most of them are greater than 0.99. Figure 2 provides visual illustrations of the comparisons of unbalanced value and balanced value for four types of pollutants.

**Comparison with provincial official statistics.** Among all provinces in China, only Zhejiang province reports sectoral level environmental data. We collect sectoral level data for four types of pollutants from Zhejiang Statistical Yearbook on Environment and Zhejiang Natural Resources and Statistical Yearbook on Environment and coordinate all industrial sectors according to Table 4. Figure 3 provides visual illustrations of the comparisons of official value and our value for four types of pollutants. We can see that observations for all pollutants are well fitted and OLS results in Fig. 3 are highly consistent to those provided in Fig. 2.

**Comparison with national aggregation data.** The national aggregation data of four types of pollutants have been revised due to the correction of micro-level firm data. Figure 4 presents the comparisons of original and revised national aggregation data. Results show that the revised national aggregation data are smoother than the original data over the time. Abnormal data such as national COD removal data in 2002 and 2012 have been corrected. Almost all discharge and removal data of four types of pollutants have been adjusted downward for all years. Removal data have higher reduction rates than discharge data.

**Code availability**

In this study, we use the General Algebraic Modeling System (GAMS) to conduct the cross-entropy estimation, and MINOS solver is used to conduct the nonlinear optimization tasks. All codes used for analysis are available in the public GitHub repository: https://github.com/qianhaoqi/China-Industrial-Environmental-Database.

Received: 22 October 2021; Accepted: 27 April 2022;
Published online: 01 June 2022

**References**

1. Shen, J. A simultaneous estimation of Environmental Kuznets Curve: Evidence from China. *China Econ. Rev.* **17**, 383–394 (2006).
2. Brajer, V., Mead, R. W. & Xiao, F. Searching for an Environmental Kuznets Curve in China’s air pollution. *China Econ. Rev.* **22**, 383–397 (2011).
3. Lan, J., Kakinaka, M. & Huang, X. Foreign Direct Investment, Human Capital and Environmental Pollution in China. *Environ. Resour. Econ.* **51**, 255–275 (2012).
4. Zhang, C. et al. An integrated specification for the nexus of water pollution and economic growth in China: Panel cointegration, long-run causality and environmental Kuznets curve. *Sci. Total Environ.* **609**, 319–328 (2017).
5. Xu, T. Investigating Environmental Kuznets Curve in China–Aggregation bias and policy implications. *Energy Policy* **114**, 315–322 (2018).
6. Liu, Q. & Wang, Q. How China achieved its 11th Five-Year Plan emissions reduction target: A structural decomposition analysis of industrial SO2 and chemical oxygen demand. *Sci. Total Environ.* **574**, 1104–1116 (2017).
7. Hang, Y., Wang, Q., Wang, Y., Su, B. & Zhou, D. Industrial SO2 emissions treatment in China: A temporal–spatial whole process decomposition analysis. *J. Environ. Manage.* **243**, 419–434 (2019).
8. Shao, L., Yu, X. & Feng, C. Evaluating the eco-efficiency of China’s industrial sectors: A two-stage network data envelopment analysis. *J. Environ. Manage.* **247**, 551–560 (2019).
9. Qian, Y., Cao, H. & Huang, S. Decoupling and decomposition analysis of industrial sulfur dioxide emissions from the industrial economy in 30 Chinese provinces. *J. Environ. Manage.* **260**, 110142 (2020).
10. Li, X., Wu, X. & Zhang, F. A method for analyzing pollution control policies: Application to SO2 emissions in China. *Energy Econ.* **49**, 431–459 (2015).
11. Shi, X. & Xu, Z. Environmental regulation and firm exports: Evidence from the eleven Five-Year Plan in China. *J. Environ. Econ. Manage.* **89**, 187–200 (2018).
12. Jiang, L., Lin, C. & Lin, P. The determinants of pollution levels: Firm-level evidence from Chinese manufacturing. *J. Comp. Econ.* **42**, 118–142 (2014).
13. Kahn, M. E., Li, P. & Zhao, D. Water Pollution Progress at Borders: The Role of Changes in China’s Political Promotion Incentives. *Am. Econ. J.-Econ. Policy* **7**, 223–242 (2015).
14. Wang, C., Wu, J. & Zhang, B. Environmental regulation, emissions and productivity: Evidence from Chinese COD-emitting manufacturers. *J. Environ. Econ. Manage.* **92**, 54–73 (2018).
15. Zhang, B., Chen, X. & Guo, H. Does central supervision enhance local environmental enforcement? Quasi-experimental evidence from China. *J. Public Econ.* **164**, 70–90 (2018).
16. Greenstone, M. & Hanna, R. Environmental Regulations, Air and Water Pollution, and Infant Mortality in India. *Am. Econ. Rev.* **104**, 3038–3072 (2014).
17. Gehrisz, M. The effect of low emission zones on air pollution and infant health. *J. Environ. Econ. Manage.* **83**, 121–144 (2017).
18. Zhu, J., Fan, Y., Deng, X. & Xue, L. Low-carbon innovation induced by emissions trading in China. *Nat. Commun.* **10**, 4088 (2019).
19. Maamoun, N. The Kyoto protocol: Empirical evidence of a hidden success. *J. Environ. Econ. Manage.* **95**, 227–256 (2019).
20. Andersson, J. I. Carbon Taxes and CO2 Emissions: Sweden as a Case Study. *Am. Econ. J.-Econ. Policy* **11**, 1–30 (2019).
21. Bi, C., Luo, Y., Ding, J. & Liang, L. Environmental performance analysis of Chinese industry from a slacks-based perspective. *Ann. Oper. Res.* **228**, 65–80 (2015).
22. Wu, J., Song, M. & Yang, L. Advances in energy and environmental issues in China: theory, models, and applications. *Ann. Oper. Res.* **228**, 1–8 (2015).
23. Zhang, Z. & Ye, J. Decomposition of environmental total factor productivity growth using hyperbolic distance functions: A panel data analysis for China. *Energy Econ.* **47**, 87–97 (2015).
24. Yu, J., Zhu, Q., Yin, P. & Song, M. Measuring energy and environmental performance for regions in China by using DEA-based Malmquist indices. *Oper. Res. Rev.* **17**, 715–735 (2017).
25. Hawkins, J., Ma, C., Schilizzi, S. & Zhang, F. Promises and pitfalls in environmentally extended input–output analysis for China: a survey of the literature. *Energy Econ.* **48**, 81–88 (2015).
26. Mi, Z. et al. Chinese CO2 emission flows have reversed since the global financial crisis. *Nat. Commun.* **8**, 1–10 (2017).
Acknowledgements
This work was supported by the National Key Research and Development Program of China (2020YFA0608600), the National Natural Science Foundation of China (71925010, 72121002, 71703027, 71904125), Shanghai Talent Development Fund (2021098).

Author contributions
H.Q. and L.W. conceived the study. H.Q. and F.R. performed analysis. All authors (H.Q., F.R., Y.G., R.M., W.W. and L.W.) interpreted the data. H.Q. prepared the manuscript. H.Q. and L.W. revised the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-022-01362-x.

Correspondence and requests for materials should be addressed to H.Q. or L.W.

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

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

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2022