Resources, Conservation & Recycling 186 (2022) 106525

Contents lists available at ScienceDirect

Resources, Conservation & Recycling

journal homepage: www.elsevier.com/locate/resconrec

Full length article

Advancing UN Comtrade for physical trade flow analysis: Addressing the issue of missing values

Zhihe Zhang [a], Zhihan Jiang [b,c], Chuke Chen [b,c,d], Xu Zhang [a], Heming Wang [a,*], Nan Li [b,c,d], Peng Wang [b,c,d], Chao Zhang [e,f], Fengmei Ma [b,c], Yuanyi Huang [b,g], Jianchuan Qi [h], Wei-Qiang Chen [b,c,d,**]

[a] State Environmental Protection Key Laboratory of Eco-Industry, Northeastern University, Shenyang 110819, China
[b] Key Lab of Urban Environment and Health, Institute of Urban Environment, Chinese Academy of Sciences, Xiamen 361021, China
[c] Xiamen Key Lab of Urban Metabolism, Xiamen 361021, China
[d] University of Chinese Academy of Sciences, Beijing 100049, China
[e] School of Economics and Management, Tongji University, Shanghai 200092, China
[f] UN Environment-Tongji Institute of Environment for Sustainable Development, Shanghai 200092, China
[g] College of Civil and Transportation Engineering, Shenzhen University, Shenzhen 518060, China
[h] Ganjiang Innovation Academy, Chinese Academy of Sciences, Ganzhou 341119, China

ARTICLE INFO

Keywords:
UN Comtrade
Physical trade
Missing physical value
Material flow analysis database

ABSTRACT

Trade contributes to the redistribution of resources among countries and regions. One of the most widely used data sources is the United Nations Commodity Trade Statistics Database (UN Comtrade). Nevertheless, data issues still limit its validity, trustworthiness, and use. A critical issue is the lack of commodity weight information. It relies heavily on data quality to determine the global market's suppliers and consumers. Thus, trade needs reliable methods for filling in the missing physical values. Using statistical approaches, we estimate missing physical values for commodities, countries/areas, and years. The impact of handled data on countries and commodities varies considerably; for example, South Africa's net weight rose by 117% and clocks' and watches' by 63% (HS0, 1988–2019), compared with their original data. The directions of net trade flows for 10594 records have been reversed. Finally, the bilateral asymmetry problem improved. Overall, this paper introduces a novel approach for improving data accuracy.

1. Introduction

Trade has long been a subject of great interest in various fields (Chen et al. 2019). In recent years, there has been increasing interest in evaluating the impact of trade on people's lives and the planet (Dalin et al. 2017; Zhang et al. 2017). Trade plays a crucial role in redistributing resources and wealth among countries and regions (Xu et al. 2020b; Yang et al. 2020). For example, trade is vital to circular economy development (Wang et al. 2020b). Over the last few decades, resource trade has resulted in a growing share of global resources extraction, reaching 24.8% in 2017 (Figure A1). Therefore, it is necessary to quantify the material flows accompanied by trade, which involves diverse processes.

The United Nations Commodity Trade Information Database (UN Comtrade), established in the early 1960s, is one of the most extensive and accurate international trade statistics databases. For more than 50 years, it has supplied a plethora of trade information to policymakers, business communities, academic institutions, and the general public (Comtrade 2019). It stores standardized annual, and monthly trade statistics supplied by countries/areas and reflects detailed international commodity flows between partners, accounting for up to 99% of global merchandise trade (Comtrade 2019). Many kinds of research, including our earlier ones, have shown, however, that missing values in UN Comtrade pose statistical issues that can lead to considerable trade misunderstanding (Espinoza and Soulier 2016; Nakajima et al. 2018; Shi et al. 2021) and the severity of the situation worsens as the proportion of missing values rises. This problem exists for all commodities, countries, and years (Table A1). The lack of data will result in

* Corresponding author.
** Corresponding author.
E-mail addresses: wangheming2006@gmail.com (H. Wang), wqchen@iue.ac.cn (W.-Q. Chen).

https://doi.org/10.1016/j.resconrec.2022.106525
Received 15 March 2022; Received in revised form 2 July 2022; Accepted 3 July 2022
Available online 13 July 2022

underestimating material flows and environmental influence. However, it will also cause net flow reversals (e.g., shifting from net importers to net exporters). There is thus a pressing need to address the problems associated with missing values in the UN Comtrade database.

In recent years, the proportion of missing values (namely missing monetary values, missing physical values, and both missing, see Figure A2) has increased. The missing weight data is the most common, and in this study, we focus on missing physical values. The following reasons may cause missing physical values: 1) Little focus on trade weight. Custom reports of various countries mainly focus on money rather than weights; 2) Wrong unit conversion (Brewer et al. 2020a; Brewer et al. 2020b). Some commodities are not reported in kilograms (kg, for example, natural gas). It may result in errors when converting their unit to kg; 3) The lack of unit conversion. Custom data are given in "quantity"; however, UN Comtrade data are released in "net weight". UN Comtrade may overlook some data when filling out the "net weight", resulting in "false data missing".

Previous studies have used the global average price of a specific commodity or the linear regression method (Dittrich and Bringezu 2010; Dittrich et al. 2012; Farhan 2015). In addition, UN Statistics Division (UNSD) estimates missing data using the median and unit prices (United Nations Statistics Division September,2017). These methods, however, ignore the varieties between commodities, reporters, and years, limiting their usefulness to the UN Comtrade database. For instance, these methods assume that the price of a car is constant worldwide, which is inconsistent with reality. Furthermore, these methods are inapplicable to custom data reported in kg because trade weight can be determined simply from "quantity".

This paper is the third one in this series on addressing data quality issues of the UN Comtrade database. Our first paper presents the status quo, causes, existing solutions, and challenges of data quality issues (Chen et al. 2022). The second paper establishes an improved framework to identify outliers (Jiang et al. 2022). This third paper aims to develop a framework to handle missing physical values in UN Comtrade for all commodities from 1988-to 2019. We also quantify the data quality improvements to examine the influence of missing physical values. The rest of this paper is structured as follows. The second section summarizes the primary methods used in this study. The third section presents the results and a brief critique of the findings. The fourth section compares these methods and discusses the limitations of this research. The fifth section highlights the main conclusions of this study.

2. Methods

2.1. The classification of data and data miss

UN Comtrade data includes statistics from the original unit of measure (as indicated in Table 1). Missing values include missing monetary value, missing physical value, and both missing. The missing monetary values are primarily the result of omissions; however, the reasons for missing physical values are more complicated (already introduced in the first section). This study only shows the results of missing physical values. With the same framework, missing monetary values can be handled similarly to missing weights. When the trade value is missing, the trade value can be calculated by combining the trade weight data with the price calculated from the trade data of other countries. Although we can locate them, the records (less than 0.02%) will be erased from the database if the missing values remain after processing. This is mainly because we cannot estimate the missing values by using any method or bilateral records.

2.2. The model of dealing with missing physical values

We developed a two-step framework. First, the original data reported by customs in kg are filtered, and "net weight" data not recorded in UN Comtrade trade are filled with the original "quantity" data reported by official customs, namely quantity(qty) in the raw UN Comtrade data. Second, the missing physical values are handled (Fig. 1). In particular, outliers (i.e., the data record whose unit price is unusually high or low) in these data were identified in our previous work of this series. These identified outliers are then corrected with estimated values using this framework. The data we used in this study are the UN Comtrade data during 1988–2019 for 5037 commodities based on the Harmonized System version 0 (HS0) classification.

Table 1
The comparison of different units.

Code	Unit	Description	Missing records	Missing rate (missing records / total records, %)	Trade value proportion (%)
1	-	No quantity	9911,476	34.4	8.9
2	m^2	Area in square meters	701,127	26.3	0.3
3	1000 kWh	Electrical energy in thousands of kilowatt-hours	10,355	92.7	0.2
4	m	Length in meters	148,997	62.7	0.03
5	U	Number of items	9131,917	14.1	24.7
6	2u	Number of pairs	301,239	12.1	0.7
7	l	Volume in liters	134,758	4.4	1.3
8	kg	Weight in kilograms	2297,516	0.9	63.3
9	1000u	Thousands of items	4135	20.5	0.009
10	U(jeu/ pack)	Number of packages	2450	7.5	0.002
11	12u	Dozens of items	9234	100.0	0.001
12	m^3	Volume in cubic meters	70,327	10.1	0.4
13	carat	Weight in carats	57,289	44.4	0.4

The general process of our framework is as follows. Firstly, the type of missing value needs to be determined. If the trade value with the United States Dollars (USD) unit is missing, we go directly to the next step of training models. If the weight is missing, we need to determine whether the original unit is kg or not. The data whose original unit is kg (reported by customs) can be directly obtained from the 'quantity'. If the data cannot be obtained from 'quantity', we will follow the processing framework to handle the missing values. Secondly, the data with no missing values are selected based on the missing data's attributes (i.e., year, reporter, and partner), of which 70% are used as the training dataset for model training, and 30% are used as the test dataset for model testing. Existing studies show that 70%–30% of dataset combinations work well for large samples to perform cross-validations (Nguyen et al. 2021). Specifically, we first select non-missing data that meet the conditions based on some features of the missing data (such as year, country, and commodity). Moreover, we randomly divided the non-missing data satisfying the missing data attributes into 70% and 30% datasets with 100 different combinations (James et al. 2013). Thus, we will assume that 30% of the data is missing and compare the estimate with the actual. Then we choose a more appropriate method for the missing data. Finally, we use the selected methods for estimating.

2.3. The seven specific methods and their differences

We selected seven methods based on previous studies, reality (trade prices are frequently related to partners) and missing rate. For example, if there are missing physical values on China's imports of Japan's cars, the following two factors should be considered: 1) the price of China's

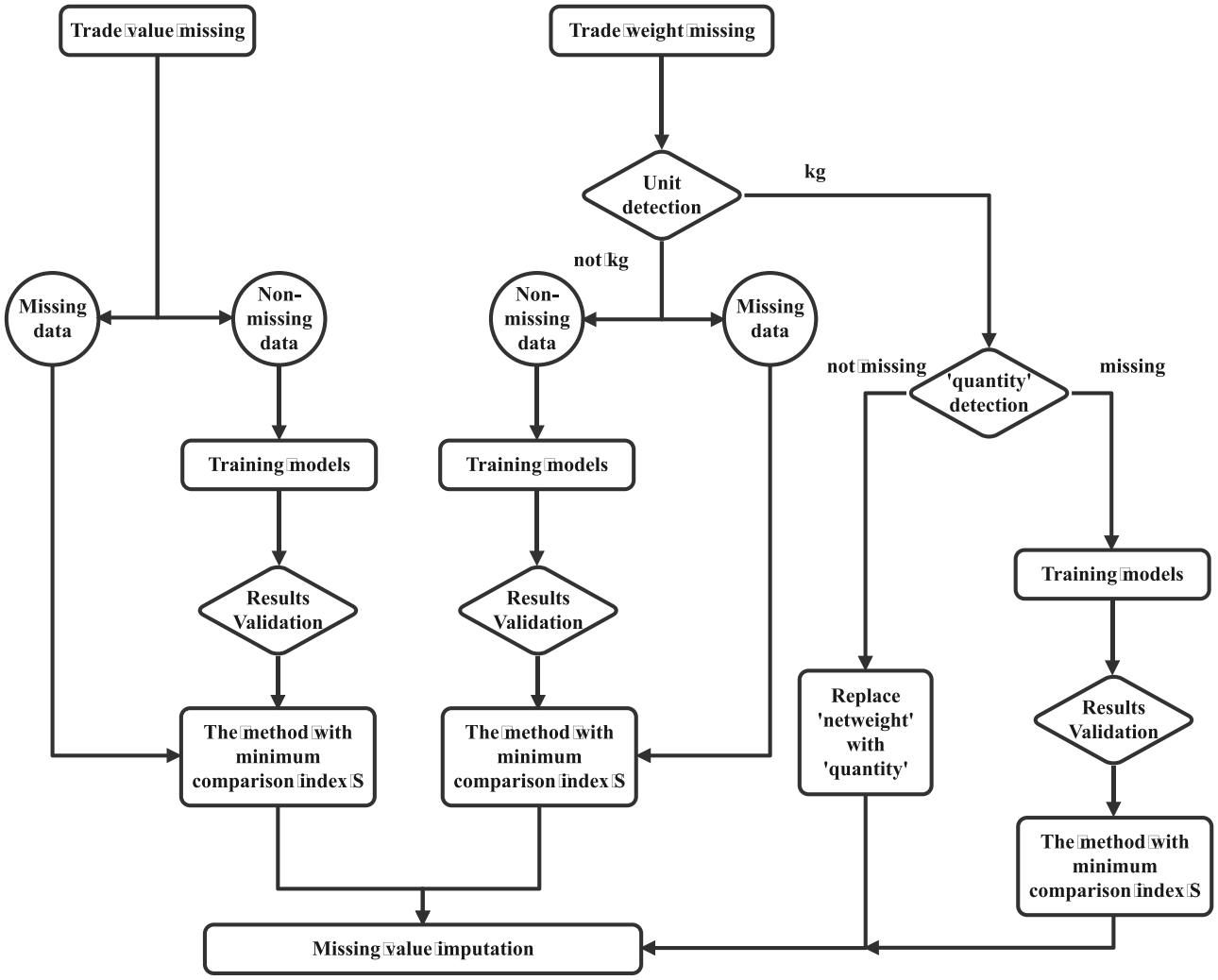

Fig. 1. Framework by which missing values are estimated.

traded cars; 2) the price of Japan's traded cars. Then, based on the missing rate of the reporter's data (China in this example), three classical methods can be chosen: median price, unit price, and average unit price. Suppose the reporter's data is substantially missing. The three methods listed above can be used for model training by using the partner data (Japan in this example). Finally, if none of the above methods performed well, the global unit price of the car was used to fill in the missing physical values. Comparisons of applicability between methods will be discussed in Section 4.

The following is an overview of the seven methods. The model training will not use cross-year data. For example, if some data (e.g., trade weight) in 2010 is missing, then only the data in 2010 will be used. This paper uses unit price, average unit price, and median unit price. In a specific sample, the unit price is the trade value divided by the trade weight; the average unit price is the sum of unit prices divided by the sample size; the median unit price is the median value calculated from unit prices. The net weight value will be determined for a missing value using reporter c, year t, commodity m, and partner i_0:

1. Average unit price of the reporter. First, the average unit price of a reporter in year t is calculated. Next, the weight value is estimated by the trade value divided by the price. This model can be easily calculated by:

$$P_{ctm} = \frac{1}{n_{ctm}} \sum_{i \neq i_0} \frac{V_{ctmi_0}}{W_{ctmi_0}} \tag{1}$$

$$W_{cti_0} = \frac{V_{ctmi_0}}{P_{ctm}} \tag{2}$$

Where

W_{ctmi} is the weight value for reporter c, year t, commodity m, and partner i;

P_{ctm} is the average unit price of the reporter c for commodity m in year t;

n_{ctm} is the number of partners reported by reporter c for commodity m in year t;

V_{ctmi} is the trade value reported by c for commodity m in year t with partner i.

2. Average unit price of partners. Instead of using the reporter's data, this method uses the partner's average unit price for estimating.

$$P_{i_0tm} = \frac{1}{n_{i_0tm}} \sum_{j \neq c} \frac{V_{i_0tmj}}{W_{i_0tmj}} \tag{3}$$

$$W_{cti} = \frac{V_{ctmi_0}}{P_{i_0tm}} \tag{4}$$

The notation here is the same as the formula above.

3. Linear regression model with the data reported by country c in year t and commodity m. The formula is given by:

$$W_{ctm} = \beta_1 V_{ctm} + \in_{ctm} \tag{5}$$

Where

β_1 is the parameter that needs to be estimated from the model;

W_{ctm} is the net weight reported by c for commodity m in year t;

V_{ctm} is the trade value reported by c for commodity m in year t;

$\in_{ctm}$ is the mistake function of this situation, which obeys a normal distribution.

4. Linear regression model with all the data in year t and commodity m. The formula is given by:

$$W_t = \beta_1 V_t + \in_{cti} \tag{6}$$

Where

β_1 is the parameter that needs to be estimated from the model;

W_t is the total net weight in year t;

V_t is the total trade value in year t;

$\in_{cti}$ is the mistake function of this situation, which obeys a normal distribution.

5. Calculation of the median of the unit price of reporter c in year t and commodity m, followed by using the trade value of partner i_0 to calculate the net weight value. For example, if data are missing for a specific commodity exported from China to other countries, and the unit prices of the non-missing values are \$1.8/kg, \$2.7/kg, and \$3.9/kg, then \$2.7/kg is the appropriate value.

6. Calculation of the median of the unit price of reporter c in year t and commodity m, followed by using the trade value of partner i_0 to calculate the net weight value. For example, if there are missing physical values for a specific commodity exported from China to a country, and the unit prices of the commodity reported in this country are \$1.1/kg, \$2.8/kg, and \$3.3/kg, then \$2.8/kg is the appropriate value.

7. Calculation of the average unit price in year t and commodity m, followed using the trade value of partner i_0 to calculate the net weight. Suppose there are missing data related to a commodity exported by China to other countries. The world trade unit price for this commodity is \$1/kg, \$3/kg, and \$3/kg, then (1 + 3 + 3) / 3 = \$2.33/kg is the appropriate value.

Comparison index. To evaluate the applicability of the seven methods, the following comparison index (Ong et al. 2013) is proposed to compare the effectiveness of the different methods:

$$S_j = \frac{1}{100}\sum_{1}^{100}\frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \widehat{y}_i|}{var(y)} \tag{7}$$

Where

S_j is the performance of the j^{th} method;

n is the number of records in the validation set;

y_i is the i^{th} weight value;

$\widehat{y}_i$ is the estimated weight values;

$var(y)$ is the variance of weight values in the validation set.

Normalized comparison index. The min-max normalized comparison index allows for comparing different methods' results (Mazziotta and Pareto 2013). For clarity, the comparison index S of each method is normalized by the following formula:

$$S_{new} = \frac{S - S_{min}}{S_{max} - S_{min}} \tag{8}$$

Where

S is the old value;

S_{new} is the new value from the normalized results;

S_{min} is the minimum value in the dataset;

S_{max} is the maximum value in the dataset.

***w* index**. Comparing original and processed data weights is the objective of this indicator. The closer the indicator is to 1, the less impact missing values have on product or country trade data. The formula is as follows:

$$w = \frac{Wh_i}{Wo_i} \tag{9}$$

Where

Wh_i is the trade weight of the Commodity i after processing or the trade weight in country i after processing;

Wo_i is the original trade weight of the Commodity i or the original trade weight in country i.

3. Results

In the Material Flow Analysis (MFA) research, the trade records and the trade weight are very important factors. Trade records can characterize the network architecture developed by countries using graph theory. Trade weight can reflect the complex changes in material flow between countries. Firstly, this section shows the distribution of the missing values by time, country, and commodity, which demonstrates the change in the trade records compared to discarding missing physical values directly. Secondly, this section shows the overall trade weight changes between original data and clean data. Finally, the specific results are described in order of the different objectives of the MFA studies. Due to the similar patterns of the 6-digit HS codes under the 2-digit codes, this study summarizes the 6-digit codes' results into the 2-digit codes. The cleaned dataset (2-, 4-, and 6-digit) can be accessed via www.macycle.org/improved-un-comtrade-data/.

3.1. Missing value distribution

Most of the missing physical values occur before 2000 and after 2006. In Fig. 2, the most noticeable pattern is the dramatic drop in the number of missing records and the missing rate in 2000. The number of missing records and the missing rate were 1.4 million and 15.4%, respectively, in 1999. After a dramatic drop in 2000, there was a gradual increase. The number of missing records climbed from 0.2 million in 2000 to 0.9 million in 2018, with a missing rate of 6.2% in 2019. This is due to a significant reduction in missing values for commodities in item and meter units (Figure A3). Before 2000, there was a significant quantity of missing values for commodities with these two units. However, after 2000, there was a significant reduction. Other goods in kilogram units contributed significantly to missing physical values after 2000. This also indicates that the actual values for replacing the missing physical values may be obtained easily from the original custom data. After 2000, the number of missing records for some commodities with non-kg units dropped dramatically, as seen in Figure A3.

Missing physical values pose severe problems in American countries, and the total physical trade of African countries is susceptible to missing physical values (Fig. 2). Missing physical values are the most in the USA and Canada, with 3 million and 2.2 million records, respectively, representing 24.8% and 34.1% of their respective data totals. Aside from these two countries, Singapore and China are missing 1.4 million and 1 million records, respectively, representing 26.6% and 11.4% of their total data. Turkmenistan and Canada have more significant data missing rates than the rest of the world (51.7% and 33.2%). Most of these countries usually export goods, particularly resource-related items, and weight data for these commodities are usually lacking. Natural gas is the most common commodity. Its original units are cubic meters; UN Comtrade did not convert these to weight on time may explain why there are so many missing records for this commodity. These issues are simple to resolve. The overall trade weight in these countries was dramatically raised once the data was handled. Lesotho and South Africa were the two countries that improved the most, with gains of 165.4% and 116.6%, respectively. Countries with modest trade volumes accounted for the majority of the countries with minor increases. Table A2 provides detailed information.

Missing physical values primarily exist in high-value-added commodities (Table A3), such as Commodity 85 (electrical machinery and equipment and parts thereof; sound recorders and reproducers; television image and sound recorders and reproducers, parts and accessories of such articles), Commodity 84 (nuclear reactors, boilers, machinery

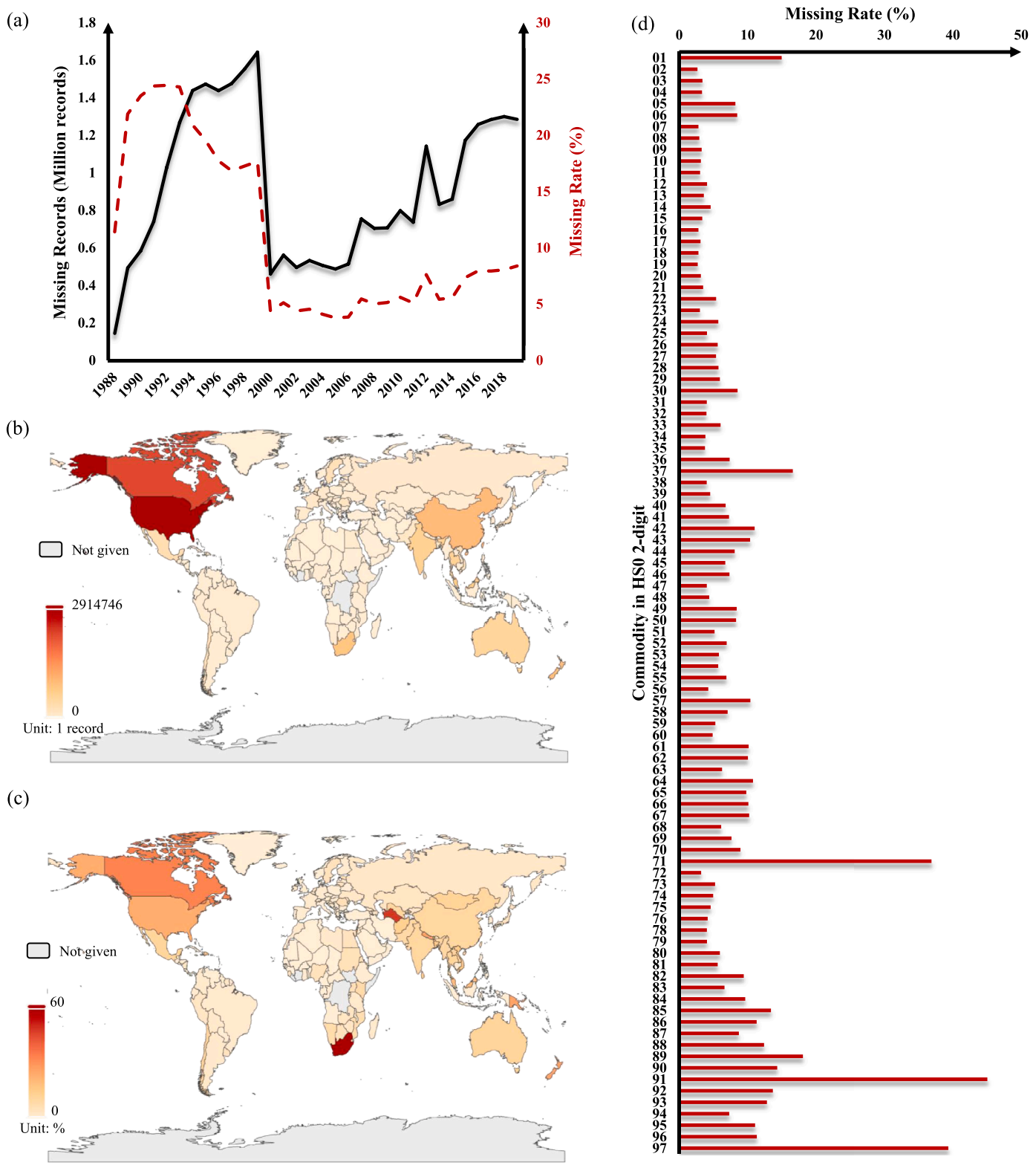

Fig. 2. Changes in the number of missing records and the missing rate during 1988–2019. (a) missing records and rate over the years; (b) missing records of reporters; (c) missing rate of reporters; (d) missing rate of commodities.

and mechanical appliances; parts thereof), and Commodity 91 (clocks and watches and parts thereof). Commodities in Commodity 85 (electrical machinery and equipment and parts thereof; sound recorders and reproducers; television image and sound recorders and reproducers, parts and accessories of such articles) are critical for economic development, with a trade value proportion of 14.6% but a missing rate of more than 10%. This type of commodity research may lead to certain misunderstandings. For example, the principal net exporter of machinery and equipment could be a net importer. Misinterpretations like these do not help policymakers since they prevent the desired regulatory

effect.

The lack of data on Commodity 84 (nuclear reactors, boilers, machinery and mechanical appliances; parts thereof) can be attributed to two factors. First, the units of most Commodity 84 (nuclear reactors, boilers, machinery and mechanical appliances; parts thereof) are not kg; hence, unit conversion issues may arise. Second, it is usually a critical strategic component. Some governments may require that the quantity of these items is kept confidential and that only a value quantity be recorded rather than the exact quantity. The weight of Commodity 91 (clocks and watches and parts thereof) is not usually mentioned due to its tiny size. Because clock compositions may vary, a proper weight estimation technique for determining their weight does not exist; as a result, the missing rate for these commodities is considerable. The total trading weight of these items increased dramatically due to our processing.

Missing physical values are more common in commodities that are not reported in kilograms (Table A3). The weight of various daily commodities, such as meat (Commodity 02), coffee beans (Commodity 09), lead (Commodity 10), and some agricultural commodities, did not considerably rise after processing. This is because the bulk trade of these commodities is recorded in unit kg, and the problem of missing trade weight data may affect only a tiny portion of the trade. Clocks (Commodity 91), pearls (Commodity 71), and art collections (Commodity 97) were the three items whose trading weight increased significantly following data processing. These commodities are usually light in weight, yet they have been demonstrated in some life cycle assessments to have significant hidden carbon emissions and contribute to resource consumption issues. This suggests that a failure to analyze this data properly could lead to a significant underestimating of the impact of these commodities on the environment and resource use.

After addressing the missing physical values, the trade weights of all countries and commodities improved to varying degrees (Fig. 3). Most countries' *w* are nearly 1 on a national scale; however, South Africa stands out since its ratio is about 3. This also indicates that if data are not handled, the material flows of South Africa's trade would be severely underestimated. Commodity 91 (Clocks and watches, and parts thereof) and Commodity 97 (Works of art; collectors' items and antiquities) are the commodities most vulnerable to missing physical values from a commodity perspective.

Interestingly, we found that some countries had many missing records but a low *w* (near to 1), and vice versa. The USA, for example, has the most missing records, and the *w* is only 1.08. This has happened in countries such as Canada, China, and Australia. South Africa, on the other hand, is the polar opposite. This suggests that, for countries with relatively substantial missing records but a small *w*, the uncertainty in these studies should rarely be assigned to missing physical values and that other alternative causes (such as model parameters) should be thoroughly examined. However, in countries where the proportion of missing physical values is low but *w* is high, relevant studies should consider the potential influence of missing physical values. Table A2 displays the entire set of results. The processing of missing physical values can improve the data quality of commodities, such as large, heavy, and expensive equipment; however, the data quality increase brought about by the processing of missing physical values for light, and sensitive commodities (such as Commodity 91- Clocks and watches, and parts thereof) is more visible. As a result of these findings, the use of UN Comtrade data to track the flow and inventory of commodities or commodity-related materials/substances may be subject to significant uncertainty in some circumstances, as a small number of missing physical values may result in large undervalued. Details are shown in Table A3.

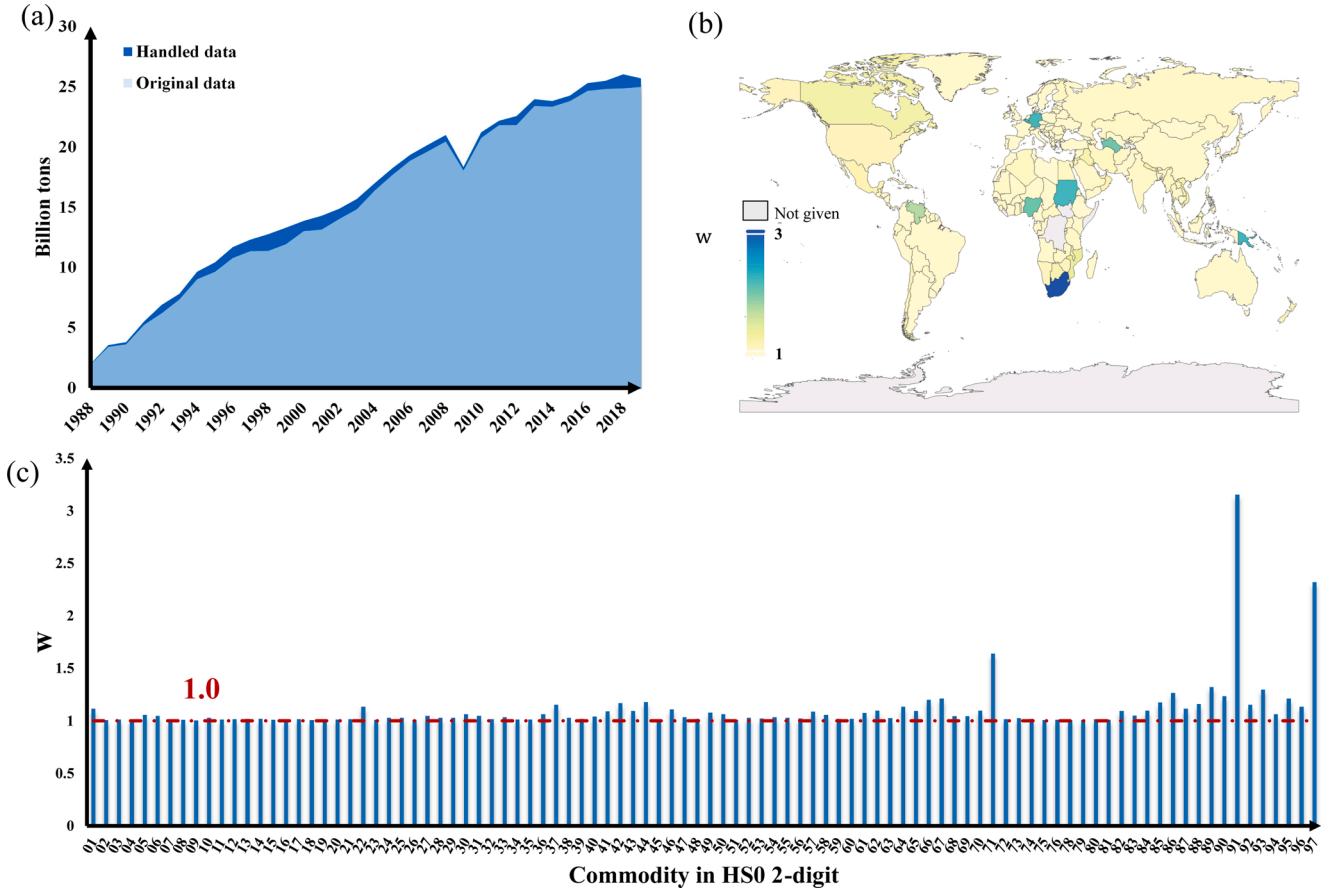

Fig. 3. Overall data improvement. a) over the years; b) of reporters (1998–2019); c) of commodities (1998–2019).

3.2. Effects of addressing the missing physical value issue

Material flow analysis (Chen et al. 2020; Fischer-Kowalski et al. 2011) and life cycle assessment (Dai et al. 2020; Ma et al. 2021) are typical methods for examining embodied resource flows in the trade process. However, due to the serious issue of missing physical values, many countries' trade data may be considerably underestimated. For example, suppose missing physical values are resolved. In that case, Venezuelan trade weight data can be increased by 61.3%. These underestimates usually influence foreign commodities associated with many carbon emissions and resource usage (Tukker et al. 2018). The database presented in this study can provide higher-quality trade data for these countries, and more trustworthy and accurate evaluations of direct and indirect material flows in the trade process, which can help policymakers establish valid policies. These advances significantly impact the major trading countries in the southern hemisphere and some industrialized countries. The database proposed in this study can be used to analyze the trade of these countries.

Another often-used approach to studying commodity flows between countries is the complex network analysis, whose results are also sensitive to missing physical values (Chen et al. 2018; Xu et al. 2020a). In the complicated network, the trading country is the node, and the trade flow is the edge (Wang et al. 2020a). The presence of missing physical values causes several misunderstandings in the network. The number of missing records corresponds to the number of missing edges, with electronic equipment and nuclear reactor equipment being the most affected commodities (Table A3). The missing edges have exceeded 100,000 edges in the last decade, which can have a detrimental impact on the analysis of the network aggregation coefficient, network resilience, and dynamic propagation of network risk of these two commodities and may lead to false findings.

The physical trade balance (PTB) indicates whether a country is a net exporter or importer of resources (Infante-Amate and Krausmann 2019; Schandl et al. 2018) We checked the trade records for all commodities, all years, and all countries/areas (Fig. 4). Many records showed PTB reversal, and there is also a similar situation with monetary trade, but that issue is not discussed here. For example, suppose that one country is regarded as a net importer of cars. After dealing with missing physical values based on the original data, it becomes a net exporter of cars based on the handled data. We define this as "reversal". In 2019, there were 1492 PTB reversals, 642 from a net importer to a net exporter and 850 from a net exporter to a net importer.

At the national level, there were multiple reversals of net imports and exports (Table A10). For example, in Mexico, a total of 2115 records were reversed, with 2102 of these being switches from net exporters to net importers, the majority of which were for Commodities 25 (225 records, Salt; sulfur; earths, stone; plastering materials, lime and cement), 39 (186 records, Plastics and articles thereof), 73 (163 records, Iron or steel articles), 84 (202 records, Nuclear reactors, boilers, machinery and mechanical appliances; parts thereof), and 85 (202 records, Nuclear reactors, boilers, machinery and mechanical appliances; parts thereof). As a result, missing physical values are expected to significantly impact trade studies of these commodities in Mexico. In addition, 1190 records in the USA were reversed; 1148 of these records were switches from net importers to net exporters, the majority of which were for Commodity 84 (296 records, Nuclear reactors, boilers, machinery and mechanical appliances; parts thereof), Commodity 85 (151 records, Nuclear reactors, boilers, machinery and mechanical appliances; parts thereof), and Commodity 90 (296 records, Optical, photographic, cinematographic, measuring, checking, medical or surgical instruments and apparatus; parts and accessories). The results for New Zealand were similar to those for the USA, indicating that missing physical values are likely to impact trade studies in these countries significantly. Such data flaws must be considered in future investigations. Table A11 shows the results at the commodity level, which are identical to the results for countries. Commodity 85 (1125 records, Nuclear reactors, boilers, machinery and mechanical appliances; parts thereof), Commodity 84 (1271 records, Nuclear reactors, boilers, machinery and mechanical appliances; parts thereof), Commodity 90 (592 records, Optical, photographic, cinematographic, measuring, checking, medical or surgical instruments and apparatus; parts and accessories), and Commodity 29 (Organic chemicals) had the most reversals (753 records).

The issue of bilateral trade asymmetry, which means that the total imports and exports are not equal, is prevalent in trade data. Regardless of whether the problem is caused by outliers or missing values, the bilateral asymmetry issue can be improved to some extent (Fig. 5). For example, the difference between global total import volume and total export volume before 2016 has been drastically improved after data processing. However, the effect in 2017–2019 is not significant. Although the bilateral trade asymmetry is not examined in this research, the methods of handling the bilateral trade asymmetry after filling in the missing values are worth discussing. In some years, the import and export trends diverge, but the two trends become nearly identical once the missing values are filled in. However, there are still some years when the import and export trends diverge. It will be discussed in more detail in the fourth study of this series.

Commodity 91 (clocks, watches, and parts thereof) is an excellent example of a commodity with a high missing rate (Figure A4). Prior to

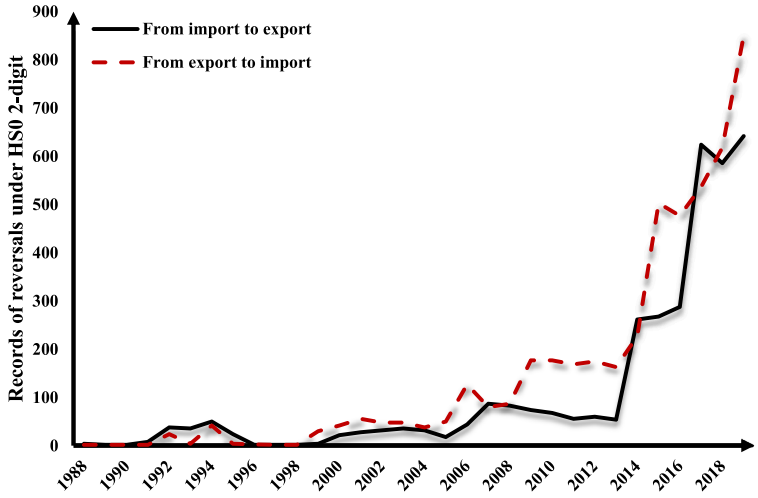

Fig. 4. The records in which net imports and net exports changed during 1988–2019. Note: "From import to export" means that in the original data, a country is a net importer of a commodity, but in the handled data is a net exporter of the same commodity in that year.

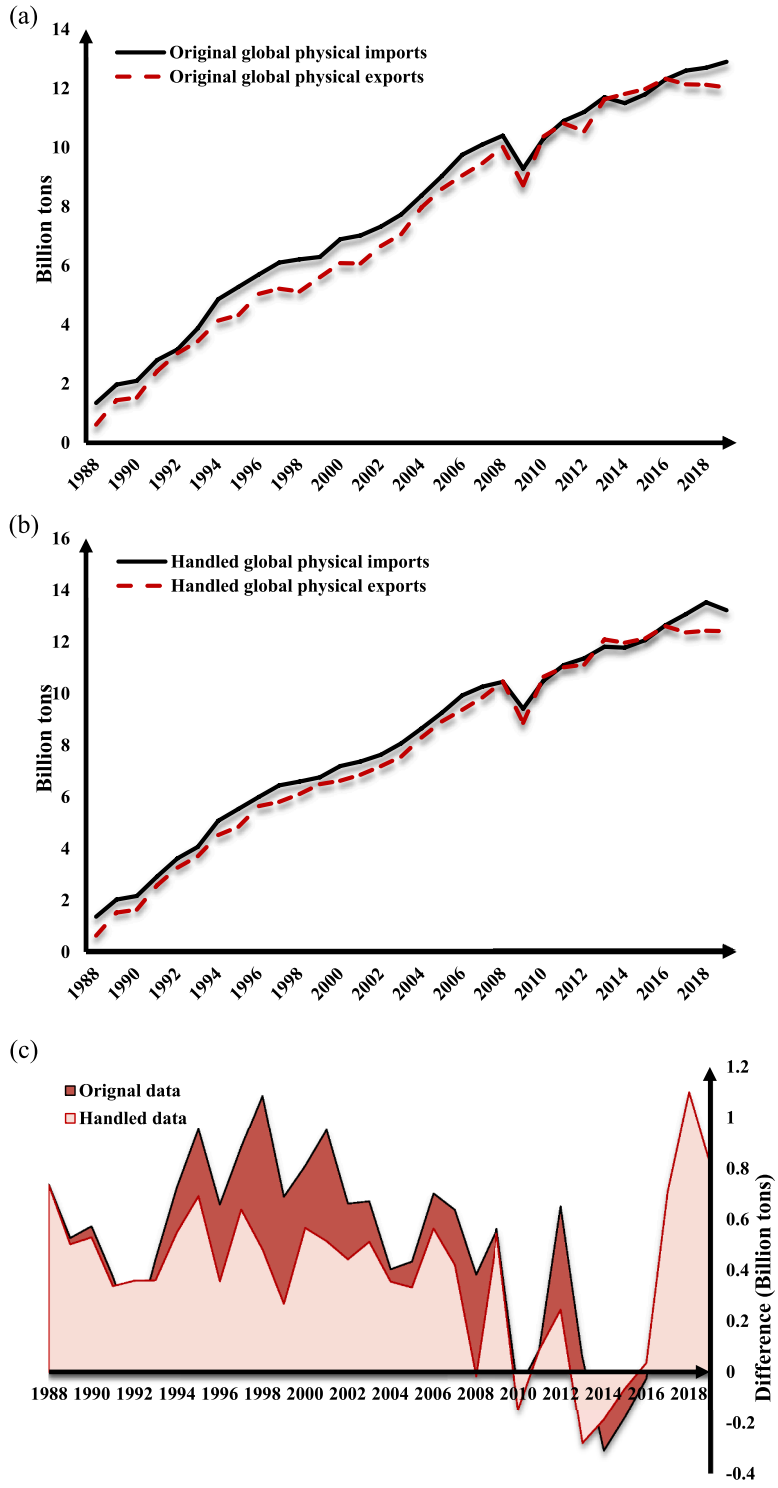

Fig. 5. Comparison of (a) original and (b) handled physical flow data, as well as (c) their difference. Note: outliers were excluded from the original data; difference = global physical imports – global physical exports.

the year 2000, the trade weight data were virtually always absent since UN Comtrade did not estimate the weight of this commodity. After 2000, the missing rate dropped significantly, remaining between 20% to 30% for the next decade. After data processing, almost all the weights increased significantly (Table A4). Many countries have seen their net trade flows shift from net exports to net imports, which has intensified in recent years. There were 46 reversals in 2019, 23 records (for example, vehicles traded in China in 2000) that moved from net imports to net exports, and 23 records that changed from net exports to net imports (all from the commodities under the 91 chapter) (Table A5). Australia, Germany, Switzerland, Liechtenstein, and Brazil were among the countries that saw these reversals (Table A6). This also suggests that the quality of these data, particularly data for the world's major watch parts, is likely to affect Commodity 91 (clocks, watches, and parts thereof) trade studies in these countries. When exporting countries become importing countries, the results can be drastically different.

According to Section 3.1, another representative commodity with large records and high missing rates is Commodity 85 (electrical machinery and equipment, and parts thereof; sound recorders and reproducers; television image and sound recorders and reproducers, and components and accessories of such commodities). Mechanical equipment plays a significant role in a country's modernization. As a result, the quality of statistics for electronic equipment is critical for determining the extent to which countries rely on electronic equipment imports. Even though the missing rate of Commodity 85 was only 12% in 2019, it grew after 2000. (Figure A5). The trade weight of this commodity grew by 13.7% in 2019 after our treatment (Table A7). There were even more records of the shift from net imports to exports of Commodity 85 (Table A8), with 164 in 2019; the shift to net imports from net exports is more widely. Most of the reversals (221 records) occurred in Brazil (Table A9), with 218 examples switching from net exports to net imports. This suggests that many of the results based on electronic equipment research in Brazil should be reevaluated and that future studies should consider similar data flaws. These shifts are in line with Brazil's reliance on imported electrical equipment. Brazil exports many agricultural commodities and natural resources like soybeans and iron ore. As a result, putting the approach outlined in this study into practice can provide new and more precise insights into electronic equipment material flows in Brazil.

4. Method comparison

Under most circumstances, one method is challenging to adapt to all countries and commodities. As a result, we compare the method performance by commodities and countries. All method performance results are available on figshare (Zhang et al. 2022).

The method selection of different commodities in the same country is different based on the performance evaluation results. Here we take China as an example since this pattern can be identified in other countries. As shown in Fig. 6, Method 1 works well when the number of samples is large, and the missing rate is low (see Table A3 for other commodities). However, it performs very poorly for high missing rate

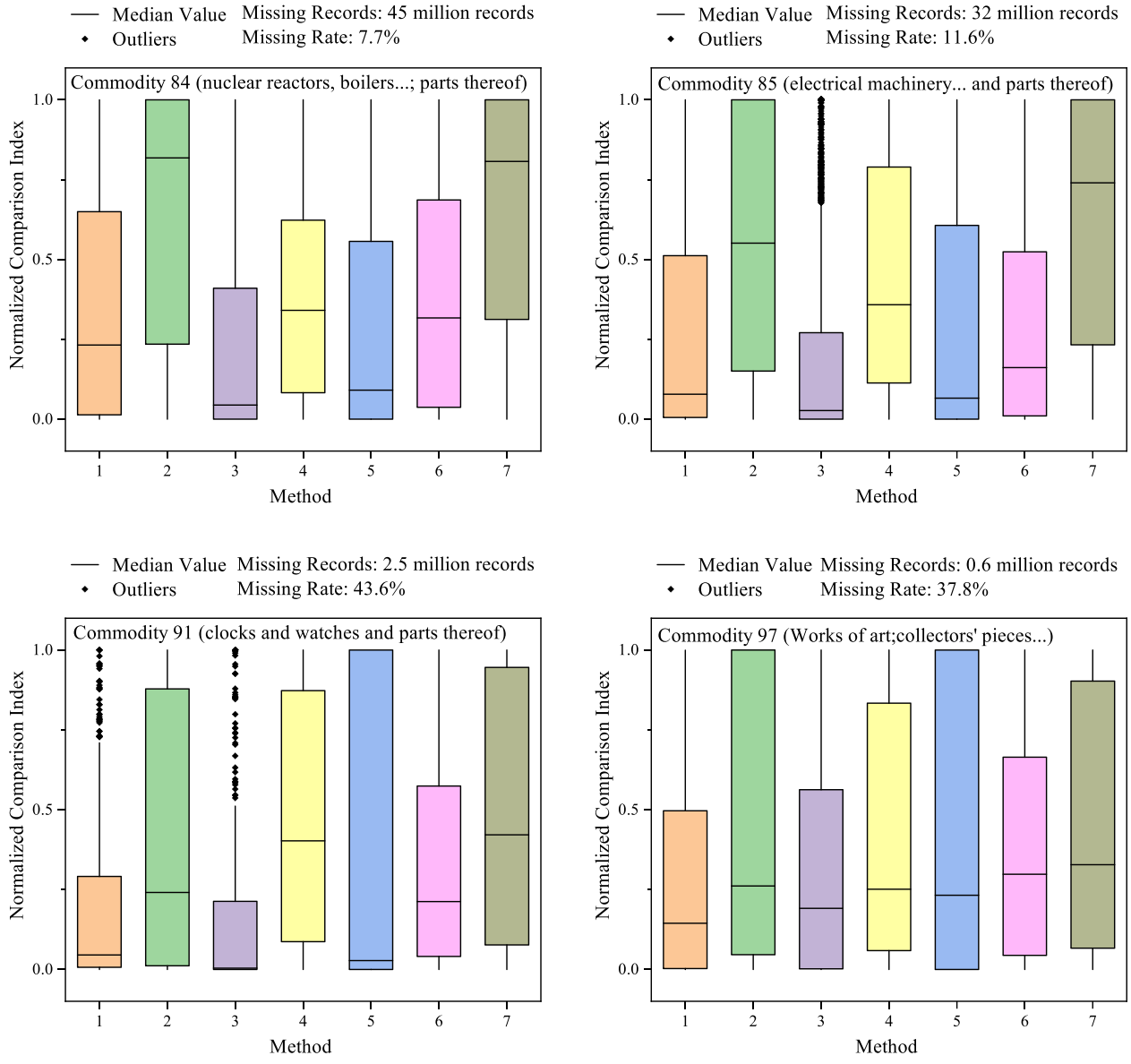

Fig. 6. Performance comparison of different methods for selected commodities in China in all years. Note: We prefer the method with fewer outliers and a low median value; the outlier in the figure is not the outliers in the UN Comtrade database.

commodities (e.g., Commodity 91- clocks and watches and parts thereof, Fig. 6). Conversely, the suitable commodities for Method 2 are different from Method 1, i.e., Method 2 works well in Commodity 91 (clocks and watches and parts thereof) but works poorly in Commodity 85 (electrical machinery and equipment and parts thereof; sound recorders and reproducers; television image and sound recorders and reproducers, parts and accessories of such articles). When the sample size is sufficient, and the missing rate is low, Method 3 performs well; however, as the missing rate increases, Method 3 becomes unstable. The accuracy rate decreases as the number of records declines. Method 4 performs better when the sample size is sufficient. If the training set data are not extensive or there are too many missing values, their performance is relatively poor. Methods 5 and 6 perform well in a variety of situations. Method 7 performs well in situations with more missing values in reporters because Method 7 only considers the prices of trading products from partner countries. Notably, Commodity 97 (Works of art; collectors' pieces and antiques) shows a different situation. When the sample size is small and the missing rate is high, the difference between the seven methods' performance becomes small.

The method selection of the same commodity in the different countries is also different. Here we take Commodity 84 (nuclear reactors, boilers, machinery and mechanical appliances; parts thereof) to illustrate the similar patterns for other commodities. China, the US, Canada, and South Africa are selected for illustration because they have more missing physical values and higher missing rates, as introduced in Section 3. Fig. 7 shows that the performance of the identical commodity in the seven methods differs substantially among countries. Methods 4 and 7 work poorly in Canada for Commodity 84 (nuclear reactors, boilers, machinery and mechanical appliances; parts thereof). This may be due to the poor performance of methods 4 and 7 in the face of samples with more missing physical values. In Fig. 7, the performance of the seven methods in China is not characterized by many outliers, showing that many methods have high applicability in China, among which Method 3 is the best. South Africa and China are in a similar situation. The situation is similar in the USA and Canada. Overall, these findings indicate that using different methods for different commodities or countries is

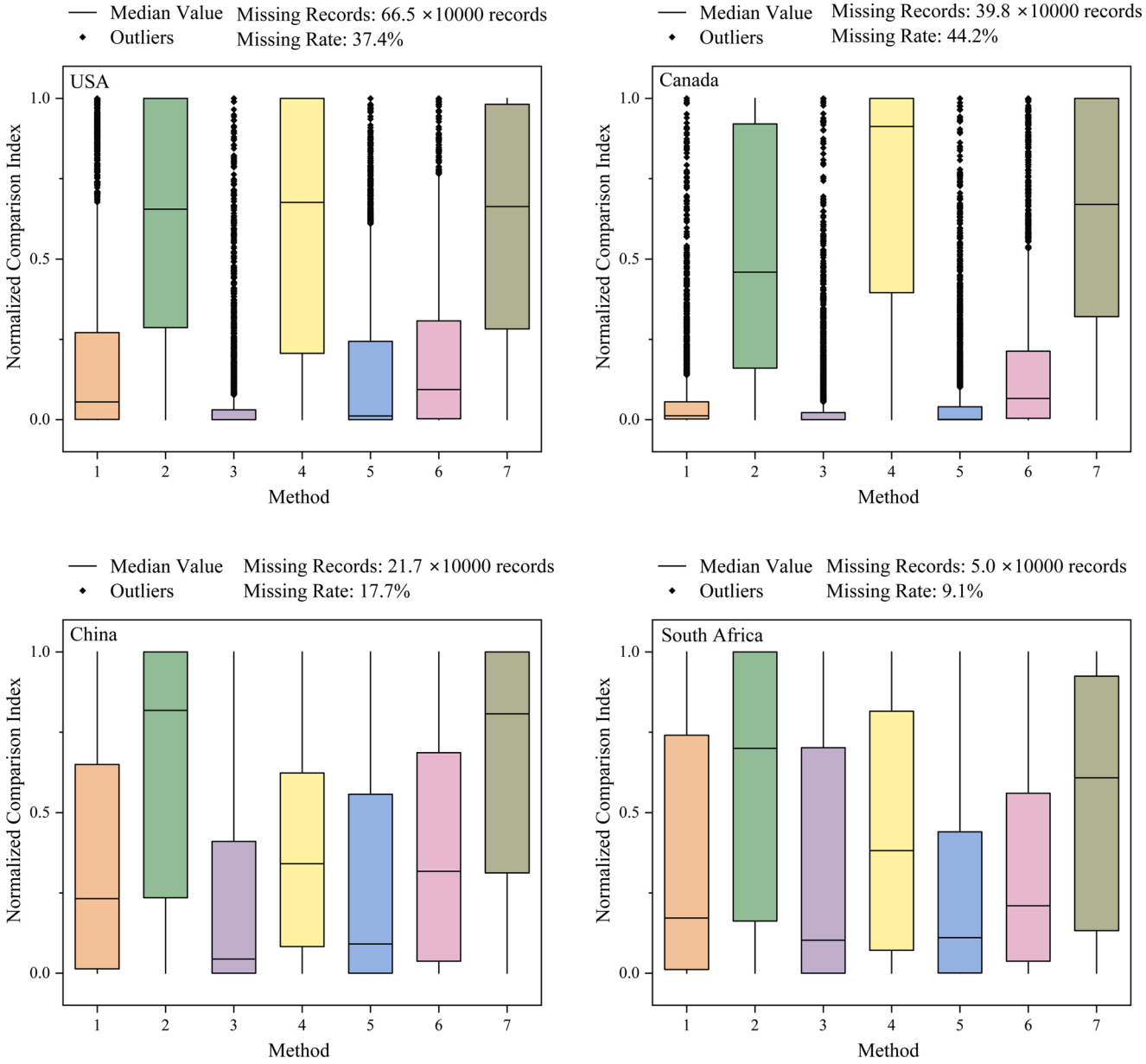

Fig. 7. Performance comparison of different methods for commodity 84 in selected countries in all years. Note: We prefer the method with fewer outliers and a low median value; the outlier in the figure is not the outliers in the UN Comtrade database.

fair. It also demonstrates that the framework we offer is helpful to the data from UN Comtrade.

To better understand these, we calculated the frequency of the optimal method. According to findings (Figure A6) from different years, Method 4 can fix roughly 66.2% of missing records. There are still many missing records that are difficult to correct. It also suggests that, in comparison to the usual method, our framework is better suited to cleaning UN Comtrade data. Method 4 is often seen in large numbers in Commodity 84 (Nuclear reactors, boilers, machinery, and mechanical appliances; parts thereof), Commodity 85 (electrical machinery and equipment and parts thereof; sound recorders and reproducers; television image and sound recorders and reproducers, parts and accessories of such articles), and Commodity 97 (Works of art; collectors' pieces and antiques). According to our findings, these commodities account for 38.4% of the data for which Method 4 was utilized to fill missing physical values (Table A12). The preceding strategy (only use one method to deal with all the missing values), on the other hand, is usually not the ideal method for missing records of other commodities. This only adds to the evidence that the procedure proposed in this study is preferable.

In summary, the application of the methods can be briefly summarized in Table 2.

Table 2
Comparison of different methods. Note: R means the missing rate and V represents missing values.

Method	Reporter	Partner
1	High R, low V	High R, high V
2	High R, high V	High R, low V
3	Low R, low V	High R, low V
4	High R, low V	High R, low V
5	Low R, high V	High R, high V
6	High R, high V	Low R, high V
7	High R, high V	Low R, low V

4.1. Comparison with existing databases

There are no comparable databases other than the official UN Comtrade raw data we used for processing for verification purposes. Developed by improving the UN Comtrade database, the existing database solely provides monetary trade information (e.g., BACI and OEC). However, our database advances the physical trade flow analysis by supplementing the current physical information.

4.2. Limitations

Methods selection is exclusively dependent on the input data's unique characteristics. We do not differentiate between imports and exports, either. Different missing value assessment methodologies applied to the same country and in different years might produce different results. For example, a country may not have a price advantage for cars. However, the price may be the same as in industrialized countries after data processing. Based on UN Comtrade, we performed in-depth optimizations that considered the variability of different countries and commodities in different years. However, because UN Comtrade refreshes the data almost daily (UN Comtrade: International Trade Statistics 2021), the handled data was usually out of date, which could be a source of error in the results, affecting our following estimation method. Their estimating procedure is typically dependent on a commodity's set conversion coefficient; as a result, the conversion coefficient's error is often substantial.

5. Conclusions

This paper presented a framework consisting of seven methods that can be used to deal with the missing values of all commodities in the UN Comtrade database. One of the main advantages of our framework is that it is based on the estimation of the data distribution, which can better distinguish the heterogeneity of all commodities in different countries and different years, which increases the accuracy of the estimation. These estimation procedures significantly improved the data quality of the physical values of global trade in UN Comtrade, which can increase the reliability of studies concerning the estimation of trade elasticity, material flow analysis, life cycle assessment analysis, complex network analysis, and others.

This study showed that the proportion of missing values had increased yearly from 1988 to 2019. The missing values are mainly concentrated in the trade data of Canada, Mexico, the USA, New Zealand, and electronic equipment, nuclear reactor parts, clocks, and pearls have the most missing values. Overall, our results allow us to draw the following four conclusions:

(1) Almost every reporter and every commodity have missing values, which can be attributed to the fact that the original units provided by the reporters are not in kg, or the UNSD did not estimate it.
(2) The missing records are significant, greatly influencing the trade flow indicators. For example, some commodities' net imports and exports in some countries showed several reversals after processing, such as Brazil. At the same time, missing values also affect the bilateral asymmetry.
(3) Our testing and processing methods are effective for trade data in UN Comtrade; the net weight of some countries and commodities increases significantly after processing the data.
(4) The data processing framework of UN Comtrade in this study can help researchers and policymakers design more effective policies. For example, the clean data can help to reduce the misjudgment of the role of net exporters due to the existence of missing physical values.

CRediT authorship contribution statement

Zhihe Zhang: Resources, Investigation, Methodology, Validation, Formal analysis, Writing – original draft. **Zhihan Jiang:** Investigation, Methodology. **Chuke Chen:** Resources, Formal analysis, Visualization, Writing – review & editing. **Xu Zhang:** Writing – review & editing. **Heming Wang:** Conceptualization, Supervision, Writing – review & editing, Funding acquisition, Methodology, Formal analysis. **Nan Li:** Conceptualization, Supervision, Writing – review & editing. **Peng Wang:** Methodology, Writing – review & editing. **Chao Zhang:** Writing – review & editing. **Fengmei Ma:** Writing – review & editing. **Yuanyi Huang:** Methodology, Validation, Data curation. **Jianchuan Qi:** Writing – review & editing. **Wei-Qiang Chen:** Conceptualization, Supervision, Writing – review & editing, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

Data will be made available on request.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (No.41871204, No. 71961147003, No. 52170184, and No.52070034).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.resconrec.2022.106525.

References

Brewer, T.D.; Abbott, D.; Lal, N.; Sharp, M.; am Thow; Andrew, N.L. (2020a): A method for cleaning trade data for regional analysis: the Pacific food trade database version 1 (1995-2016). Pacific Community working paper X.

Brewer, T.D.; Andrew, N.L.; Sharp, M.K.; am Thow; Kottage, H.; Jones, S. (2020b): A method for cleaning trade data for regional analysis: the Pacific food trade database (version 2, 1995–2018). Pacific Community working paper.* A comprehensive version of this research.

Chen, B., Li, J.S., Wu, X.F., Han, M.Y., Zeng, L., Li, Z., Chen, G.Q., 2018. Global energy flows embodied in international trade: a combination of environmentally extended input–output analysis and complex network analysis. Appl. Energy 210, 98–107. https://doi.org/10.1016/j.apenergy.2017.10.113.

Chen, Chuke, Jiang, Zhihan, Li, Nan, Wang, Heming, Wang, Peng, Zhang, Zhihe, et al., 2022. Advancing Un Comtrade for physical trade flow analysis: review of data quality issues and solutions. SSRN J. https://doi.org/10.2139/ssrn.4060058.

Chen, Wei-Qiang, Ciacci, Luca, Sun, Ning-Ning, Yoshioka, Toshiaki, 2020. Sustainable cycles and management of plastics: a brief review of RCR publications in 2019 and early 2020. Resour., Conserv. Recycl. 159, 104822 https://doi.org/10.1016/j.resconrec.2020.104822.

Chen, Wei-Qiang, Ma, Zi-Jie, Pauliuk, Stefan, Wang, Tao, 2019. Physical and monetary methods for estimating the hidden trade of materials. Resources 8 (2), 89. https://doi.org/10.3390/resources8020089.

Comtrade, U.N. (2019): UN Comtrade. In *United Nations Commodity Trade Statistics Database*.

Dai, Tao, Yang, Yi, Lee, Ross, Fleischer, Amy S., Wemhoff, Aaron P, 2020. Life cycle environmental impacts of food away from home and mitigation strategies-a review. J.Environ. Manag. 265 (PART 2), 110471 https://doi.org/10.1016/j.jenvman.2020.110471.

Dalin, Carole, Wada, Yoshihide, Kastner, Thomas, Puma, Michael J, 2017. Groundwater depletion embedded in international food trade. Nature 543 (7647), 700–704.

Dittrich, Monika;, Bringezu, Stefan, 2010. The physical dimension of international trade: part 1: direct global flows between 1962 and 2005. Ecol. Econ. 69 (9), 1838–1847.

Dittrich, Monika, Bringezu, Stefan, Schütz, Helmut, 2012. The physical dimension of international trade, part 2: indirect global resource flows between 1962 and 2005. Ecol. Econ. 79, 32–43.

Espinoza, Luis A.Tercero, Soulier, Marcel, 2016. An examination of copper contained in international trade flows. Min. Econ. 29 (2), 47–56.

Farhan, Javed, 2015. Overview of missing physical commodity trade data and its imputation using data augmentation. Transp. Res. C: Emerg. Technol. 54, 1–14.

Fischer-Kowalski, M., Krausmann, F., Giljum, S., Lutter, S., Mayer, A., Bringezu, S., et al., 2011. Methodology and indicators of economy-wide material flow accounting. J. Ind. Ecol. 15 (6), 855–876. https://doi.org/10.1111/j.1530-9290.2011.00366.x.

Infante-Amate, Juan, Krausmann, Fridolin, 2019. Trade, ecologically unequal exchange and colonial legacy: the case of France and its former colonies (1962–2015). Ecol. Econ. 156, 98–109.

James, Gareth, Witten, Daniela, Hastie, Trevor, Tibshirani, Robert, 2013. An Introduction to Statistical Learning. Springer.

Jiang, Zhihan, Chen, Chuke, Li, Nan, Wang, Heming, Wang, Peng, Zhang, Chao, et al., 2022. Advancing Un Comtrade for physical trade flow analysis: addressing the issue of outliers. SSRN J. https://doi.org/10.2139/ssrn.4060057.

Ma, Zijie, Yang, Yi, Chen, Wei-Qiang, Wang, Peng, Wang, Chao, Zhang, Chao, Gan, Jianbang, 2021. Material flow patterns of the global waste paper trade and potential impacts of China's import ban. Environ. Sci. Technol. 55 (13), 8492–8501. https://doi.org/10.1021/acs.est.1c00642.

Mazziotta, Matteo;, Pareto, Adriano, 2013. Methods for constructing composite indices: one for all or all for one. Riv. Ital. Econ. Demogr. Stat. 67 (2), 67–80.

Nakajima, Kenichi, Daigo, Ichiro, Nansai, Keisuke, Matsubae, Kazuyo, Takayanagi, Wataru, Tomita, Makoto, Matsuno, Yasunari, 2018. Global distribution of material consumption: nickel, copper, and iron. Resour., Conserv. Recycl. 133, 369–374.

Nguyen, Quang Hung, Ly, Hai-Bang, Ho, Lanh Si, Al-Ansari, Nadhir, van Le, Hiep, van Tran, Quan, et al., 2021. Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. Math. Prob. Eng. 2021, 1–15. https://doi.org/10.1155/2021/4832864.

Ong, Ghim Ping, Farhan, Javed, Chin, Anthony Theng Heng, 2013. Stochastic imputation of missing physical commodity trade information using monetary trade data. Transp. Res. Record 2354 (1), 122–132.

Schandl, Heinz;, Fischer-Kowalski, Marina;, West, James;, Giljum, Stefan;, Dittrich, Monika;, Eisenmenger, Nina, et al., 2018. Global material flows and resource productivity: forty years of evidence. J. Ind. Ecol. 22 (4), 827–838. https://doi.org/10.1111/jiec.12626.

Shi, Jiujie, Zhang, Chao, Chen, Wei-Qiang, 2021. The expansion and shrinkage of the international trade network of plastic wastes affected by China's waste management policies. Sustain. Prod. Consump. 25, 187–197.

Tukker, Arnold, Koning, Arjan de, Owen, Anne, Lutter, Stephan, Bruckner, Martin, Giljum, Stefan, et al., 2018. Towards robust, authoritative assessments of environmental impacts embodied in trade: current state and recommendations. J. Ind. Ecol. 22 (3), 585–598.

UN Comtrade: International Trade Statistics (2021): Data availability. Available online at https://comtrade.un.org/data/da, updated on 10/18/2021, checked on 10/18/2021.

Wang, Chao, Zhao, Longfeng, Lim, Ming K., Chen, Wei-Qiang, Sutherland, John W, 2020a. Structure of the global plastic waste trade network and the impact of China's import ban. Resour., Conserv. Recycl. 153, 104591.

Wang, Heming, Schandl, Heinz, Wang, Xinzhe, Ma, Fengmei, Yue, Qiang, Wang, Guoqiang, et al., 2020b. Measuring progress of China's circular economy. Resour., Conserv. Recycl. 163, 105070.

Xu, Wen, Chen, Wei-Qiang, Jiang, Daqian, Zhang, Chao, Ma, Zijie, Ren, Yan, Shi, Lei, 2020a. Evolution of the global polyethylene waste trade system. Ecosyst. Health Sustain. 6 (1), 1756925.

Xu, Zhenci, Li, Yingjie, Chau, Sophia N., Dietz, Thomas, Li, Canbing, Wan, Luwen, et al., 2020b. Impacts of international trade on global sustainable development. Nature Sustain. https://doi.org/10.1038/s41893-020-0572-z.

Yang, Lan, Wang, Yutao, Wang, Ranran, Klemeš, Jiří Jaromír, Almeida, Cecília Maria Villas Bôas de, Jin, Mingzhou, et al., 2020. Environmental-social-economic footprints of consumption and trade in the Asia-Pacific region. Nat. Commun. 11 (1), 1–9.

Zhang, Qiang, Jiang, Xujia, Tong, Dan, Davis, Steven J., Zhao, Hongyan, Geng, Guannan, et al., 2017. Transboundary health impacts of transported global air pollution and international trade. Nature 543 (7647), 705–709.

Zhihe Zhang; Zhihan Jiang; Chuke Chen; Xu Zhang; Heming Wang; Nan Li et al. (2022): Methods comparison results.