Spatiotemporal Statistical Imbalance: A Long-Term Neglected Defect in UN Comtrade Dataset

Luoming Hu 1,2, Changqing Song 1,2, Sijing Ye 1,2,* and Peichao Gao 1,2

1 State Key Laboratory of Earth Surface Processes and Resource Ecology, Beijing Normal University, Beijing 100875, China; huluoming@mail.bnu.edu.cn (L.H.); songcq@bnu.edu.cn (C.S.); gaopc@bnu.edu.cn (P.G.)
2 Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China
* Correspondence: yesj@bnu.edu.cn

Abstract: The bilateral trade data provided by the United Nations International Trade Statistics Database are some of the most authoritative trade statistics and have been widely used in many research fields. Here, we propose a new form of inconsistency in its records, namely statistical imbalance, which refers to the phenomenon of inequality between the import or export trade value of a commodity category and the total value of all its subcategories. We investigated the frequency and spatial-temporal patterns of the statistical imbalances of 15 reporters (i.e., Australia, Brazil, Canada, China, France, Germany, India, the Netherlands, the Rep. of Korea, the Russian Federation, Switzerland, the United Arab Emirates, the United States of America, and Vietnam) from 1996–2016 and explored their distributional differences in commodity categories with a co-clustering algorithm. The results show that statistical imbalance is widespread with obvious clustering patterns. Trade records related to specific categories such as fossil fuels, pharmaceuticals, machinery, and unspecified commodity categories presented severe statistical imbalances, which may lead to erroneous trade research results. Since statistical imbalance is difficult to detect in studies focusing only on specific commodity categories, we suggested that researchers should prescreen the data for statistical imbalance to ensure the validity of their results.

Keywords: United Nations International Trade Statistics Database; bilateral trade; statistical imbalance; co-clustering algorithm; geographic information system

1. Introduction

In the present era of globalization, trade is an essential component of modern society, and nations have signed bilateral trade agreements to engage in various forms of economic integration. Bilateral trade data have played an increasingly important role in various research fields, such as analyzing trade competition and cooperation among different countries or tracking global ecosystem service flows. The United Nations International Trade Statistics Database (UN Comtrade) is one of the most widely used international trade databases with a high degree of authority and uniformity. The database’s records date back to 1962, and the total quantity of records exceeds 3 billion. Over 200 reporting countries provide their annual international trade statistics data detailed by commodities or service categories and partner countries. Trade records are stored according to a classification system based on the category to which the goods belong, with similar goods falling into one large category. These data are subsequently transformed into the United Nations Statistics Division standard format with consistent coding, e.g., Harmonized Commodity Description and Coding System (HS), Standard International Trade Classification (SITC), and Classification by Broad Economic Categories (BEC) and valuation in the data loading process.

UN Comtrade has made significant contributions to multiple research topics and policies, demonstrating the importance of trade records in the governance of economic
activities [1,2]. First, UN Comtrade has provided basic data for enhancing the recognition of trade systematic rules and its driving factors. For instance, Veninga et al. investigate the effects of domestic political instability on the wheat trade in Egypt [3]. Oluwatoba et al. evaluate the impact of a Free Trade Agreement (FTA) on South African agricultural trade by using the Poisson Pseudo Maximum Likelihood (PPML) specification of the gravity model [4]. Other studies investigate the effects of FTAs on European agri-food trade [5], Korean seaborne trade [6], and Latin American export diversification [7].

Second, UN Comtrade has provided practical guidance for developing measurement methods of international trade. Complex network analysis has been extensively used to reveal the structures and evolution of trade relationships and interdependencies among trade partners, which are not immediately evident in a straightforward statistical analysis of trade data [8]. For instance, Cristelli quantified export similarity by building distance matrices for products and countries based on the complex network and then determining the evolution of competitors’ communities [9]. Dong constructed wheat-trading competition networks to analyze the impact of climate change on the global trade flows of wheat and then proposed a policy framework to promote a stable and healthy wheat-trading environment [10]. Other optional methods include the gravity model [4,5,11] and the digital trade feature map method [12,13].

Third, UN Comtrade has provided support for depicting global trade patterns and changing processes. Numerous indexes have been proposed to estimate the trade relationships (e.g., comparative advantage, complementarity, similarity, and technical complexity) of specific commodities or industry chains among different countries. For instance, Zheng calculated and compared the technical sophistication index and its regional heterogeneity of new energy products and new energy industries among 30 countries [14]. Cao measured the evolution of the technical complexity of China’s export environmental goods and its position in the international industrial value chain [15]. Hao analyzed the overall characteristics of the iron ore importing competition pattern, the import competition region, and the main importing countries [16]. Xu calculated the trade competitiveness index and investigated the impact of Chinese textiles on UK imports [17].

Fourth, UN Comtrade has been used increasingly more widely as a medium for other themes, such as global ecological protection [18–21], pollution prevention [22], energy management [23], and national security [24]. Moran calculated the embodied ecological footprint of specific countries exerted inside the borders of their trading partners [25]. A similar method was also used to analyze the historical terms of trade in footprint units of key agricultural commodities traded between the US and Britain during the 19th century [21]. To enrich the conceptualization and policy discussion of global electronic waste, Lepawsky quantified the magnitude and direction of this trade between 206 territories in over 9400 reported trade transactions from 1996 to 2012 [22]. Dalin combined agricultural trade flows with province-level estimates of commodities’ virtual water content to build China’s domestic and foreign virtual water trade network and then analyzed the virtual water flow patterns as well as the corresponding water savings [26]. Meyfroidt proposed that assessments of countries’ contributions to reforestation and carbon emission reductions should integrate the geographic displacement of forest clearing across countries through trade in agricultural and forest products and calculated the percentage of net wood trade offset to the total reforested area of specific countries when both agriculture and forestry sectors are included [19]. Chen evaluated the energy security of Kazakhstan and Turkmenistan (as exporters) and Kyrgyzstan (as importers) according to correlation, diversity, and the impact of international relations using energy trade data from the UN Comtrade [23].

Despite the universality and authority in applications, UN Comtrade is inconsistent because its data sources are compiled on a different country-of-origin basis without continuous observation. A typical form of this inconsistency is referred to as “bilateral asymmetries”, which occur when the reported exports from country A to country B do not match the reported imports to country B from country A. “Bilateral asymmetries” create incomparability and detract from the usefulness of such trade data for some types
of economic analysis. Veronese et al. empirically validated the bilateral asymmetries and discussed the reasons for it in the Mediterranean partner countries [26]. Markhonko then summarized several main reasons for bilateral asymmetries, including “application of different trade systems in data compilation”, the “time lag between exports and imports”, and “imports and exports are respectively reported in CIF-type and FOB-type values” [27]. Bousey compared Canada’s and China’s bilateral trade data and analyzed major sources of asymmetry [28]. United Nations Statistics Division in 2019 discussed ways to measure, analyze, and reduce bilateral asymmetries [29]. In contrast, we find another form of inconsistency through integrity checks of merchandise trade data and refer to it as “statistical imbalance”, which can be described as the reported exports or imports of a specific commodity category not matching the summation of their all subcategories. Since data perfection has not received enough attention from researchers for a long time, few studies have focused on this form of inconsistency. This may lead to large deviations in research results if we neglect this phenomenon. To fill this gap, we sorted out and discussed the patterns, reasons and possible responses to “statistical imbalance”.

In this paper, we firstly introduced the data and methods we used in Section 2, including the details of data acquisition and processing in our automatic ETL application program, as well as spatiotemporal judgment matrix for analyzing statistical imbalances and how to construct it. Then, we introduced and discussed the co-clustering algorithm. In Section 3, we calculated the occurrence frequency of statistical imbalance and its spatiotemporal distribution pattern by clustering reporters, partners, and years. We separately showed the patterns of the statistical imbalance in annual trade volume and commodity categories organized by HS 2-digit code using the co-clustering algorithm. What is more, we analyzed the results in Section 4 and discuss the causes of the statistical imbalance and limitations of our research. Finally, in conclusion, we summarized the results and proposed several strategies for reducing statistical imbalance. Overall, the statistical imbalance is a flaw in the data. These differences range from a few dollars to billions of dollars and have a very wide distribution in the UN Comtrade database. While it is acceptable to ignore some small discrepancies, we should be wary of the threat that serious and extreme ones which may pose to the accuracy and reliability of the study in relevance. If a researcher happens to use these data without noticing its imperfection, it may lead to conclusions that are contrary to the facts or incomparable between studies. Compared to bilateral asymmetries, severe statistical imbalances are no less significant in terms of trade volume, but have long been underappreciated. We hope our research can provide references for UN Comtrade data selection and integrity checks as well as help UN Comtrade to have a more profound impact in the new era of digital governance [30,31].

2. Data and Methods
2.1. Data

Trade statistics are big data with long time series. On this basis, to explore the occurrence frequency and spatiotemporal pattern of the “statistical imbalance” phenomenon, a Python scrapy-based ETL application program has been developed to rapidly extract annual bilateral commodity trade records (unit: USD) of 15 reporters (i.e., Australia, Brazil, Canada, China, France, Germany, India, the Netherlands, the Rep. of Korea, the Russian Federation, Switzerland, the United Arab Emirates, the USA, and Vietnam) from 1996–2016. The selection of reporters has considered several factors such as geographical position, scale of foreign trade, commodity category, and social and economic development. All of these factors may have an impact on the accuracy, richness, and quality control of trade records. These countries have relatively complete and representative trade data for the period from 1996 to 2016, which can avoid errors due to uncertainty. Moreover, because the volume of UN Comtrade data is very large, analysis using all of the data is difficult to achieve, while some of the representative data are sufficient to efficiently obtain meaningful conclusions. For each reporter, its corresponding partners are set as all countries and regions except itself. All the commodity categories are organized with consistent and
nested HS (The Harmonized Commodity Description and Coding System) rules, including HS 2-digit codes, HS 4-digit codes, and HS 6-digit codes. The HS 2-digit codes correspond to 99 commodity categories. The HS 4-digit codes and HS 6-digit codes are generated from further subdivision of HS 2-digit codes and HS 4-digit codes, respectively. The HS 6-digit codes comprise approximately 5300 commodity categories.

The logical execution process of the ETL application program is listed as follows (see Figure 1):

**Figure 1.** Logical execution process of the ETL application program. (a) Automatic ETL processing system for international trade data. (b) Random check system.

Step 1. Data request URLs (Uniform Resource Locator) have been constructed automatically by adjusting reporters, partners, years, trade flow types (i.e., imports and exports), and HS codes. Then, for each URL, steps 2–4 were executed independently.

Step 2. Check if the URL has already been successfully executed: if yes, go to the next URL and restart step 2; if no, request the URL, obtain the records, and proceed to step 3.

Step 3. Check if the records are successfully acquired. If yes, load the records to the database, record the URL, and proceed to step 4; if no, rollback the task, record the error URL in the error log file, go to the next URL, and return to step 2.

Step 4. Check if all URLs have been executed: if yes, end the traversal; if no, go to the next URL and return to step 2.

The above process should be executed multiple times until no error URL has been recorded. The development and use of ETL applications have greatly improved the efficiency of data downloading, while avoiding errors that can occur due to duplication and omissions. In addition, one manual strategy was integrated into the data extraction process to check for unexpected data errors (omissions, repetition, and unreadable data) and guarantee the integrity of the data. Using the strategy, multiple records were randomly extracted to compare with the results obtained manually from the UN Comtrade official
website. A total of 44,659,097 trade records were filtered as experimental data, including 2,232,953 HS 2-digit-code records, 12,173,521 HS 4-digit-code records, and 30,252,623 HS 6-digit-code records.

2.2. Data Spatio-Temporal Judgment Matrix for “Statistical Imbalance”

In theory, for commodity trade records with a specific reporter, partner, year, and trade flow type, the total trade volume of all commodity categories organized by HS 2-digit codes should be equal to that of the commodity subcategories organized by HS 4-digit codes. However, in practice, there are probably some differences between them, which can be treated as “statistical imbalance”. The estimation of these differences in export trade statistics has been formulated as (1), for example. In the equation, \( i \) and \( j \), respectively, represent the reporter and partner, \( k \) is the year, and \( m \) and \( n \) represent specific commodity categories organized by HS 2-digit code and HS 4-digit code, respectively. \( V_{EXP,i,j,k,m} \) represents the corresponding export trade volume of commodity \( m \). \( DV_{EXP,i,j,k,m} \) represents the degree of export imbalance between reporter \( i \) and partner \( j \) in year \( k \). Furthermore, since imbalanced features express huge differences as the condition changes, \( DV_{EXP,i,j,k,m} \) has been converted to \( LG_{EXP,i,j,k,m} \) using (3). In addition, imports and exports are exactly similar.

\[
DV_{EXP,i,j,k,m} = \sum_{m \in HS2} V_{EXP,i,j,k,m} - \sum_{n \in HS4} V_{EXP,i,j,k,n} \\
LG_{EXP,i,j,k,m} = \frac{DV_{EXP,i,j,k,m}}{|DV_{EXP,i,j,k,m}|} \log_{10} \left( \frac{|DV_{EXP,i,j,k,m}| + 1}{|DV_{EXP,i,j,k,m}|} \right) 
\]

Spatiotemporal judgment matrixes were constructed to intuitively express the distribution of imbalanced features and its relation to reporter, partner and year conditions, as Table 1 show. Row names are unduplicated groups of reporters-years, and column names are lists of partners. Table elements are in accordance with \( LG_{EXP,i,j,k,m} \). On that basis, reporter-partner-year groups with conspicuous export imbalances (judgment criteria: \( |LG_{EXP,i,j,k,m}| > 4 \)) were extracted. In addition, for each group, the export imbalance \( DV_{EXP,i,j,k,m} \) of each HS 2-code-based commodity category \( m (m = 1, 2, \ldots, 99) \) is calculated by using the total exports of HS 4-code-based commodity subcategories belonging to it (i.e., \( C_{HS4} \)) minus that of its further subdivided HS 6-code-based commodity subcategories (i.e., \( C_{HS6} \)), as (3) shows. On that basis, another spatiotemporal judgment matrix (E-STJM-HS2C) was constructed by taking reporter-partner-year groups as rows and HS 2-codes as columns to present the impacts of commodity categories on export imbalance features, as Table 2 shows. Table elements \( LG_{EXP,i,j,k,m} \) are converted from \( DV_{EXP,i,j,k,m} \) through (4). Moreover, cases in imports and exports are exactly similar.

\[
DV_{EXP,i,j,k,m} = \sum_{C_{HS4}} V_{EXP,i,j,k,C_{HS4}} - \sum_{C_{HS6}} V_{EXP,i,j,k,C_{HS6}} \\
LG_{EXP,i,j,k,m} = \frac{DV_{EXP,i,j,k,m}}{DV_{EXP,i,j,k,m}} \log_{10} \left( \frac{|DV_{EXP,i,j,k,m}| + 1}{|DV_{EXP,i,j,k,m}|} \right) 
\]
2.3. Co-Clustering Algorithm

After identifying the statistical imbalance in the data, we would like to further explore whether it shows some pattern. As one of the most common methods of data mining, clustering classifies data and finds information by measuring the similarity of attributes, structures, and information within data, which can efficiently extract patterns and effective information because of its characteristic of considering data elements at a high level of abstraction [32]. Moreover, clustering is very suitable for tasks involving processing large amounts of data [33]. Different from traditional one-way clustering that only uses data objects or attributes as features to perform similarity calculations, co-clustering algorithms equally consider data objects and attributes while clustering [34] so that their results are more meaningful. To date, co-clustering has been well developed and applied in many fields and works [33], [35–37]. Considering its ability to identify and classify features quickly and efficiently for large amounts of data, the co-clustering algorithm was applied to analyze the distribution characteristics of the bilateral trade data statistical imbalance in all countries (or areas), years, and commodity categories in our study. Rows of the spatiotemporal judgment matrix proposed above have been treated as data objects (combination of reporter and year or reporter, partner, and year) and columns as attributes (partner or commodity code).

The process is based on the Bregman block average co-clustering algorithm with I-divergence [36]. The process of running the algorithm is shown in Figure 2. First, the co-clustering algorithm randomly maps data objects and attributes to different clusters and generates the co-clustered data matrix as initialization. Then, the differences between the original matrix of statistical imbalance $O_{SU}$ and the newly generated co-clustered matrix $\hat{O}_{SU}$ are determined according to I-divergence, where $D_f(O_{SU}||\hat{O}_{SU})$ represents the I-divergence between $O_{SU}$ and $\hat{O}_{SU}$. Next, the algorithm updates the mapping from data objects and attributes to the corresponding clusters through an iterative process, that is, the
algorithm assigns different combinations of reporters and years (or reporters, partners, and years) and partners (or commodity codes) to the closest set to minimize the loss until its value reaches a local minimum or falls below a predetermined threshold. Since the global optimal result is difficult to determine, this process may be repeated in multiple initial random mappings to produce the best possible co-clustering result. Finally, the rows and columns of the original matrix are exchanged, putting the rows and columns belonging to the same cluster together to make a reordered data matrix, based on the results of the co-clustering algorithm.

Figure 2. Pseudocode for block average co-clustering algorithm with I-divergence (from Wu et al., 2016).

3. Results

3.1. Spatio-Temporal “Statistical Imbalance” of Annual Trade Volume

The “statistical imbalance” of import and export volumes between different reporters and partners in different years have been estimated according to the equations mentioned above. Empirically, we grouped every two orders of magnitude of statistical imbalance. The probability distributions in different ranges are shown in Table 3. The table shows that the statistical imbalance is mainly distributed in the range of small absolute values. For instance, the $DV_{IMP_{i,j,k}}$ proportion and $DV_{EXP_{i,j,k}}$ proportion of the range $[-10^2, 10^2]$ are, respectively, 97.899% and 95.836%. However, it is worth noting that the proportions of $DV_{IMP_{i,j,k}} = 0$ or $DV_{EXP_{i,j,k}} = 0$ are both less than 65%, indicating that statistical imbalance is widespread, but in most cases, the value is small so that it has little effect on the results of related studies. In addition, $DV_{IMP_{i,j,k}}$ and $DV_{EXP_{i,j,k}}$ have wide ranges of values, and cases of $DV_{IMP_{i,j,k}}$ (or $DV_{EXP_{i,j,k}}$) > 0 and $DV_{IMP_{i,j,k}}$ (or $DV_{EXP_{i,j,k}}$) < 0 both exist. Therefore, statistical imbalance cannot be simply attributed to missing data. In some cases, the absolute values of $DV_{IMP_{i,j,k}}$ and $DV_{EXP_{i,j,k}}$ are greater than $10^6$ or even more extreme, which means that neglecting statistical imbalance may cause serious errors. Comparing exports and imports, the probability of statistical imbalance in exports is higher, and this phenomenon is more common for a severe imbalance ($[DV_{IMP_{i,j,k}} (or DV_{EXP_{i,j,k}})] > 10^6$, where the mod operator means to take the absolute value). Overall, we considered $<10^2$ to be a small difference and $>10^6$ (including millions to billions of dollars) as serious differences because, despite their relatively low frequency of occurrence, they are already likely to have a significant impact on the associated trade analysis.
### Table 3. Probability distribution of statistical imbalance in different ranges.

| Range               | Total Proportion | $DV_{\text{IMP}_{i,j,k}}$ Proportion | $DV_{\text{EXP}_{i,j,k}}$ Proportion |
|---------------------|------------------|-------------------------------------|-------------------------------------|
| $(10^6, +\infty)$   | 1.373%           | 0.739%                              | 1.986%                              |
| $(10^4, 10^6]$      | 1.196%           | 0.763%                              | 1.618%                              |
| $(10^2, 10^4]$      | 0.541%           | 0.571%                              | 0.515%                              |
| $(10, 10^2]$        | 2.364%           | 2.027%                              | 2.695%                              |
| $(0, 10]$           | 16.380%          | 15.045%                             | 17.719%                             |
| 0 (consistent)      | 59.439%          | 63.143%                             | 55.796%                             |
| $[-10, 0)$          | 16.503%          | 15.667%                             | 17.313%                             |
| $[-10^5, -10)$      | 2.122%           | 1.950%                              | 2.292%                              |
| $[-10^4, -10^5]$    | 0.044%           | 0.067%                              | 0.021%                              |
| $[-10^3, -10^4]$    | 0.007%           | 0.006%                              | 0.008%                              |
| $(-\infty, -10^3)$ | 0.032%           | 0.022%                              | 0.037%                              |

Note: $DV_{\text{IMP}_{i,j,k}}$ and $DV_{\text{EXP}_{i,j,k}}$ represent the degrees of import and export imbalance between reporter $i$ and partner $j$ in year $k$, respectively. The total number of records is 127,121, and the numbers for $DV_{\text{IMP}_{i,j,k}}$ and $DV_{\text{EXP}_{i,j,k}}$ are 62,553 and 64,568, respectively.

Spatiotemporal judgment matrixes of import imbalances and export imbalances were constructed, as shown in Figures 3–6. The missed data (No Data) represents that the reporter country has no trade records with the partner country in a given year, which is out of scope of our study. The statistical imbalance is significantly clustered by reporter. Namely, for a specific combination of reporter and year, their corresponding partners often have similar statistical imbalance characteristics. Severe statistical imbalances are mainly concentrated in a few countries and occur in similar years, including France (1996 and 1998–1999), the Rep. of Korea (1996 and 1998), Germany (1996–1999 and 2012–2016), the Netherlands (2012 and 2014–2015), India (1999), Switzerland (1999), and the United Arab Emirates (2000–2003). However, these cases only cover some of the partners, and their characteristics are inconsistent. Severe export imbalance is more frequent, but the probability of $LG_{\text{IMP}_{i,j,k}} > 10$ is higher than that of $LG_{\text{EXP}_{i,j,k}} > 10$, indicating that import imbalance has a higher probability of extreme situations. Moreover, import imbalance and export imbalance tend to have a strong correlation, but there are exceptions, such as Germany (1996–1999). In addition, the cases where there are no commodity trade records (that is, no data areas) in both matrixes mainly appear on the right side, corresponding to partners that are small regions or islands, such as HMD ( Heard Island and McDonald Islands), SGS (South Georgia and The South Sandwich Islands), and VAT (Holy See (Vatican City State)), which is reasonable. The records data of Vietnam and the United Arab Emirates from 1996–2000 is also missing.

### 3.2. Spatio-Temporal “Statistical Imbalance” Divided by HS 2-Codes

To explore the differences in statistical imbalance of reporter, partner, and year on different commodity types, $LG_{\text{IMP}_{i,j,k,m}}$ and $LG_{\text{EXP}_{i,j,k,m}}$ were calculated using (5)–(8) and were used to construct spatiotemporal judgment matrixes of the import (and export) imbalance of different HS 2-digit-code-based commodity categories (E-STJM-HS2C or I-STJM-HS2C, as shown in Table 3). The results show that the number of elements where $LG_{\text{IMP}_{i,j,k}} > 4$ is 916, and the number of elements where $LG_{\text{EXP}_{i,j,k}} > 4$ is 2300. Then, the Bregman block average co-clustering algorithm with I-divergence (BBAC_I) was applied to E-STJM-HS2C and I-STJM-HS2C to express the congregation characteristics of the statistical imbalance generated by the interaction of reporter-year-partner groups and commodity categories, as Figures 7 and 8 show.
Figure 3. Spatio-temporal judgment matrixes of import imbalance. No data indicate that the specific reporter has no record of commodity trade with the specific partner in the labeled year. The matrix contains 315 rows (15 reporters * 21 years) and 299 columns (partners). Row names are unduplicated reporter–year groups, and column names are list of partners, which are represented by ISO country code. FRA: France, DEU: Germany, NLD: the Netherlands, CHE: Switzerland, KOR: The Rep. of Korea, IND: India, and ARE: The United Arab Emirates. The color of the legend indicates the severity of the statistical imbalance. Red (blue) represents a large positive (negative) difference, and green represents a small absolute value of the difference.
Figure 4. Total import imbalance of selected countries to all partners. Countries are selected based on those groupings that exhibit significant statistical imbalances in the spatio-temporal judgment matrixes of import imbalance.

Figure 5. Spatiotemporal judgment matrixes of export imbalance. The markings and labels are the same as Figure 3.
Figure 6. Total export imbalance of selected countries to all partners. Countries are selected based on those groupings that exhibit significant statistical imbalances in the spatio-temporal judgment matrices of export imbalance.

Figure 7. Co-clustering results of the import spatiotemporal judgment matrix by commodity category.
Figure 8. Co-clustering results of the export spatiotemporal judgment matrix by commodity category.

Figure 7 divides the spatiotemporal pattern of the import imbalance into three reporter-year-partner clusters (Clusters A, B, and C) and four commodity category clusters (Clusters 1, 2, 3, and 4). Cluster A consists of elements with widespread high statistical imbalance features and covers nearly all commodity categories. Its corresponding reporter is Germany (2012–2016). As Cluster B-1 shows, partial records of Germany (2012–2016), the Netherlands (2012 and 2014–2015), and the United Arab Emirates (2003) express lower congregation characteristics of statistical imbalance; and the relevant commodity categories are mineral fuels (HS 2-digit code: 27), inorganic or organic chemicals (28–29), pharmaceutical products (30), chemical products (38), plastics and articles (39), iron or steel articles (72–73), machinery and mechanical appliances (84), and commodities not specified according to kind (99).
Figure 8 shows that the clustering structure of E-STJM-HS2C is similar to that of I-STJM-HS2C. The statistical imbalance feature distributions of Cluster A of these two judgment matrices are consistent with each other, and their corresponding reporters are both Germany (2012–2016). There are more commodity categories in Cluster B that show severe statistical imbalance, and the distribution is more clustered (compared to Figure 7). Furthermore, nearly all negative $LG_{EXP_{i,j,k,m}}$ values are gathered in category 99. The results indicate that the statistical imbalance among different reporters, partners, and years shows spatiotemporal variation with differences according to commodity category. For a certain reporter-year group, its statistical imbalance is usually clustered among several commodity categories, except for Germany 2012–2016, where statistical imbalance covers almost all commodity categories.

4. Discussion

UN Comtrade provides basic data for research in a number of areas, but trade statistics inevitably exhibit shortcomings of inconsistency due to the complexity of international trade activities and data archiving. Inconsistencies in trade statistics may trigger bias in the results of related studies, and very often such errors are attributed by researchers as being a result of the sample or the methodology, but the data were not given sufficient attention. Typically, the imperfection of trade statistics is widely recognized [38], but the impact of this imperfection on empirical results has not received much attention [39]. Furthermore, although some researchers and organizations have discussed the “bilateral asymmetries” of UN Comtrade data, analyzing the causes and giving some suggestions to reduce their impact [26–29], few studies have paid attention to the phenomenon of “statistical imbalance” proposed in this study. We think that “statistical imbalance” is more insidious than “bilateral asymmetries”. In some cases, the trade value associated with “statistical imbalance” is large, and it may be more damaging to trade studies because it is not easily detectable. In addition, many existing studies referring to imperfections in trade statistics may focus on a specific type of problem or country/region without examining and discussing the dataset as a whole. On this basis, we have studied the UN Comtrade data for multiple representative countries in a long time series and systematically analyzed the spatiotemporal pattern of the “statistical imbalance” phenomenon.

The spatiotemporal judgment matrices of the statistical imbalance in UN Comtrade show that the distribution of the statistical imbalance exhibits a clear pattern across countries, years, and product types, rather than occurring randomly. Compared to the “bilateral asymmetries” phenomenon, which is characterized by large international trading countries (developed countries) usually having higher absolute trade differences and small trading countries (less developed countries) having higher relative trade differences [38], statistical imbalance is more concentrated in space (country) and time (year). In the period covered by this study (1996–2016), serious statistical imbalance was mainly found in trading countries (e.g., Germany, France, Korea, and Switzerland) and in the early years (around 2000) and involved fossil fuels, chemicals, pharmaceuticals, and iron and steel products, as well as machinery and unspecified commodity categories. The most serious country is Germany with the category “commodities not specified by type” (99), and this is also one of the very few categories with a negative statistical difference. In fact, we think that at some point, for example, when the “commodities not specified by type” has the opposite sign of the difference from the other commodity categories, it could also be a source of statistical imbalance. In addition, there is usually some correlation between statistical imbalance in imports and exports, and they may occur in consecutive years and for similar product categories. Therefore, we recommend conducting statistical imbalance tests prior to the study, especially for high-risk countries, years, and product categories. Methodologically, researchers can first perform a test on a random selection of data and then expand the examination if a statistical imbalance is found. While most trade statistics have no or only mild statistical imbalance that has little impact, a severe and large statistical imbalance such as Germany from 2012–2016 must be given sufficient attention. If there are serious problems
with the data, measures should be considered to ensure that the results are authentic and credible, including avoiding countries, years, and commodity categories where statistical imbalance is concentrated; using data from local national sector statistics for corroboration and supplementation; or referring to other datasets and statistics to ensure the reliability of related studies. In general, statistical imbalance deserves attention and, accordingly, data prescreening is more important in cases when trade data are used as the basis for analysis at large spatial and temporal scales, for specific categories of commodities, or for quantitative comparisons with similar studies.

The “statistical imbalance” in UN Comtrade can be attributed to multiple reasons. Numerically, most (more than 90%) of the statistically imbalanced records have small differences in absolute values, and such small differences may be due to errors in the statistical process of the original data. For example, in 2009, Canada provided trade records on imports from France with 2-digit HS codes, and the total annual trade value was $4,950,577,714; however, for 4-digit HS codes, the total annual trade value was $4,950,577,701, with a difference of only $13. The $13 error came from meat products (02) and grains (10), and other product trades have single-digit differences. Furthermore, there are two reasons corresponding to the large difference in statistical imbalance: one is the loss of data items for an entire category, such as 2-digit HS code trade records missing for commodities coded as 99 in German exports to the UK in 2016 (HS-2 < HS-4); and, second, the differences in the trade value of several commodity categories and their accumulation lead to large differences in the total trade value. For example, the total trade value of France’s exports to Japan in 1997 differed by more than 100 million US dollars, mainly from the accumulation of the differences in the trade value of powdered industrial products (11), gum and other plant sap (13), miscellaneous foodstuffs (21), and other categories of commodities. Considering that the statistical imbalance does not originate from bilateral trade but from reporters’ own trade statistics, we suggest that UN Comtrade should strengthen the inspection of the data interchange process so as to control data from the source or mark the data quality so that institutions or researchers can obtain the corresponding information when using the data.

Finally, our study contains several limitations. First, due to data limitations, we are unable to fully explore the exact causes of statistical imbalance, such as whether the imbalance comes from a reporter’s own trade statistics errors or errors in data archived by UN Comtrade. Moreover, we selected only some of the reporters and did not characterize the spatiotemporal distribution of statistical imbalance globally, the summary of patterns based on only some of the countries, years, and commodity categories with significant statistical imbalances may not be perfect. Finally, we do not specify the quantitative impact of statistical imbalances in related studies of trade analysis, and this may need to be discussed and compared in the context of specific studies. Future work should focus on these aspects in order to help researchers more comprehensively understand and avoid the impacts of such factors on studies.

5. Conclusions

In this paper, we proposed a new form of inconsistency in the UN Comtrade dataset, namely, statistical imbalance. Statistical imbalance refers to the mismatch between the import or export trade value of a specific commodity category and the total value of all its subcategories. Here, we investigated the frequency of the statistical imbalance phenomenon and its spatial and temporal patterns as well as summarized its distribution differences in commodity categories using co-clustering algorithm. The results indicated that statistical imbalance is widespread in UN Comtrade statistics and that there are clear clustering patterns. For some countries, years, and commodity categories, statistical imbalances occur significantly more frequently. In general, trade statistics of trading countries in early years on fossil fuels, chemicals, iron and steel machinery, and unspecified commodity categories are at higher risk. For these high-risk countries, years, and commodity categories, researchers need to strengthen the testing of data quality when using relevant trade statistics data. For the given reporters and years, reporters’ trade with many partners may
have similar statistical imbalance characteristics, e.g., the Netherlands’ import imbalance in 2012–2015 and France’s export imbalance in 1996–1999. This feature may provide a perspective for the rapid detection of statistical imbalances. That is, for reporters with statistical imbalances with some of their trading partners, the accuracy of their trade records with other partners during similar time periods are questionable and need to be examined in a focused manner. In addition, there is a correlation of statistical imbalance between import and export records in some trade statistics, which is also a point of concern. This means that if statistical imbalance is found in the import trade records of some commodity categories during a certain time period, then, correspondingly, the export trade records are at a relatively higher risk and need to be checked more intensively, and vice versa. In fact, although most of the recorded differences are small, there are still cases of large absolute values that cannot be ignored, such as the German import records from 2012–2016, which show large discrepancies in almost all commodity categories. The most serious inconsistency appeared in Germany in 2012–2016, where statistical imbalances were detected in almost all commodity categories. Severe statistical imbalance can significantly jeopardize the perfection of trade statistics and, thus, the validity of the research results based on them. This could cause problems for scholars using this dataset for relevant research and policy makers for decision-making. Considering that statistical imbalances are usually concentrated in a few commodity groups for a given country and year as well as that the import and export imbalance is somewhat correlated, studies using such data should perform targeted prescreening as much as possible. At the same time, it is the responsibility of government statistical offices, as producers of data, to pay sufficient attention to this problem. We strongly recommend that the United Nations Statistics Division (UNSD) make statistical imbalance testing a necessary quality control component in the data archiving process, to minimize statistical imbalance and give corresponding quality markers for scholars, experts, and policy makers.

**Author Contributions:** Conceptualization, S.Y. and C.S.; methodology, S.Y.; software, L.H.; validation, L.H.; formal analysis, L.H.; investigation, P.G.; resources, C.S.; data curation, S.Y.; writing—original draft preparation, L.H. and S.Y.; writing—review and editing, S.Y.; visualization, L.H.; supervision, P.G.; project administration, C.S.; funding acquisition, C.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the second Tibetan Plateau Scientific Expedition and Research Program (STEP) [grant number 2019QZKK0608], National Natural Science Foundation of China [grant number 42171250, 41801300, 41901316].

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data presented in this study are available on request from the corresponding author.

**Acknowledgments:** We would like to thank the high performance computing support from the Center for Geodata and Analysis, Faculty of Geographical Science, Beijing Normal University [https://gda.bnu.edu.cn/ (accessed on 25 November 2021)].

**Conflicts of Interest:** The authors declare no conflict of interest.

**References**

1. Erkut, B. From digital government to digital governance: Are we there yet? *Sustainability* 2020, 12, 860. [CrossRef]
2. Kaya, T.; Sagsan, M. The Concept of ‘knowledgization’ for Creating Strategic Vision in Higher Education: A Case Study of Northern Cyprus. *Egit. Ve Bilim.* 2016, 41, 184. [CrossRef]
3. Veninga, W.; Ihle, R. Import vulnerability in the Middle East: Effects of the Arab spring on Egyptian wheat trade. *Food Secur.* 2018, 10, 183–194. [CrossRef]
4. Fadeyi, O.A.; Bahta, T.Y.; Ogundeji, A.A.; Willemse, B.J. Impacts of the SADC free trade agreement on South African agricultural trade. *Outlook Agric.* 2014, 43, 53–59. [CrossRef]
5. Matkovski, B. The effects of Foreign agri-food trade liberalization in South East Europe. *Ekon. Cas.* 2018, 9, 945–966.
6. Lee, P.T.W.; Lee, T.-C.; Yang, T.-H. Korea-ASEAN free trade agreement: The implications on seaborne trade volume and maritime logistics policy development in Korea. *J. Int. Logist. Trade* 2013, 11, 3–26. [CrossRef]

7. Dingemans, A.; Ross, C. Free trade agreements in Latin America since 1990: An evaluation of export diversification. *Cepal Rev.* 2012, 108, 27–48. [CrossRef]

8. Hidalgo, C.A.; Klinger, B.; Barabási, A.-L.; Hausmann, R. The product space conditions the development of nations. *Science* 2007, 317, 482–487. [CrossRef]

9. Cristelli, M.; Tacchella, A.; Gabrielli, A.; Pietronero, L.; Scala, A.; Caldarelli, G. Competitors’ communities and taxonomy of products according to export fluxes. *Eur. Phys. J. Spec. Top.* 2012, 212, 115–120. [CrossRef]

10. Dong, C.; Yin, Q.; Lane, K.J.; Yan, Z.; Shi, T.; Liu, Y.; Bell, M.L. Competition and transmission evolution of global food trade: A case study of wheat. *Phys. A Stat. Mech. Its Appl.* 2018, 509, 998–1008. [CrossRef]

11. Mupela, E.; Szirmai, A. Communication Costs and Trade in Sub Saharan Africa: A Gravity Approach. In Proceedings of the International Conference on e-Infrastructure and e-Services for Developing Countries, AFRICOMM 2013, Blantyre, Malawi, 25–27 November 2013; Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. Bissyandé, T., van Stam, G., Eds.; Springer: Cham, Switzerland, 2013; pp. 27–38.

12. Ye, S.; Song, C.; Cheng, C.; Shen, S.; Gao, P.; Zhang, T.; Chen, X.; Wang, Y.; Wan, C. Digital trade feature map: A new method for visualization and analysis of spatial patterns in bilateral trade. *ISPRS Int. J. Geo Inf.* 2020, 9, 363. [CrossRef]

13. Ye, S.; Cheng, C.; Song, C.; Shen, S. Visualizing bivariate local spatial autocorrelation between commodity revealed comparative advantage index of China and USA from a new space perspective. *Environ. Plan. A. Econ. Sp.* 2021, 53, 223–226. [CrossRef]

14. Zheng, H.-H.; Wang, Z.-X. Measurement and comparison of export sophistication of the new energy industry in 30 countries during 2000–2015. *Renew. Sustain. Energy Rev.* 2019, 108, 140–158. [CrossRef]

15. Cao, X.; Hanson-Rasmussen, N. Dynamic change in the export technology structure of china’s environmental goods and its international comparison. *Sustainability* 2018, 10, 3508. [CrossRef]

16. Hao, X.; An, H.; Sun, X.; Zhong, W. The import competition relationship and intensity in the international iron ore trade: From network perspective. *Resour. Policy* 2018, 57, 45–54. [CrossRef]

17. Xu, J. The role of China in the UK relative imports from three selected trading regions: The case of textile raw material industry. *Int. J. Envion. Res. Public Health* 2017, 14, 1481. [CrossRef]

18. Peters, G.P.; Minx, J.C.; Weber, C.L.; Edenhofer, O. Growth in emission transfers via international trade from 1990 to 2008. *Proc. Natl. Acad. Sci. USA* 2011, 108, 8903–8908. [CrossRef]

19. Meyfroidt, P.; Rudel, T.K.; Lambin, E.F. Forest transitions, trade, and the global displacement of land use. *Proc. Natl. Acad. Sci. USA* 2010, 107, 20917–20922. [CrossRef]

20. Dalin, C.; Hansakai, N.; Qiu, H.; Mauzeral, D.L.; Rodriguez-Iturbe, I. Water resources transfers through Chinese interprovincial and foreign food trade. *Proc. Natl. Acad. Sci. USA* 2014, 111, 9774–9779. [CrossRef]

21. Hornborg, A. Footprints in the cotton fields: The Industrial Revolution as time—Space appropriation and environmental load displacement. *Ecol. Econ.* 2006, 59, 74–81. [CrossRef]

22. Lepawsky, J. The changing geography of global trade in electronic discards: Time to rethink the e-waste problem. *Geogr. J.* 2015, 181, 147–159. [CrossRef]

23. Xiaopeng, C.; Shengkui, C.; Liang, W. Quantitative Analysis of Central Asian Countries’ Energy Security and Its Political Influence Factors. *J. Resour. Ecol.* 2018, 9, 434–443. [CrossRef]

24. Jackson, M.O.; Nei, S. Networks of military alliances, wars, and international trade. *Proc. Natl. Acad. Sci. USA* 2015, 112, 15277–15284. [CrossRef] [PubMed]

25. Moran, D.D.; Wackernagel, M.C.; Kitzes, J.A.; Heumann, B.W.; Phan, D.; Goldfinger, S.H. Trading spaces: Calculating embodied ecological footprints in international trade using a product land use matrix (PLUM). *Ecol. Econ.* 2009, 68, 1938–1951. [CrossRef]

26. Veronese, N.; Tyrman, H. MEDSTATII: Asymmetry in Foreign Trade Statistics in Mediterranean Partner Countries; Eurostat Methodologies Working Papers; Eurostat: Luxembourg City, Luxembourg, 2009.

27. Markhonko, V. Asymmetries in Official International Trade Statistics and Analysis of Globalization. In Proceedings of the International Conference on the Measurement of International Trade and Economic Globalization, Aguascalientes, Mexico, 29 September–1 October 2014.

28. Boushey, N. Comparing Canada’s and China’s Bilateral Trade Data. Available online: https://www150.statcan.gc.ca/n1/pub/13-605-x/2018001/article/54962-eng.htm (accessed on 25 November 2021).

29. UNSD (UNITED NATIONS STATISTICS DIVISION). IMTS Bilateral Asymmetries—How to Measure, Analyze, Reduce and Way Forward; United Nations Statistics Division: New York, USA, 2019.

30. Erkut, B. Product innovation and market shaping: Bridging the gap with cognitive evolutionary economics. *Indraprastha J. Manag* 2016, 4, 3–24.

31. Sharma, G.D.; Mahendra, M. Thrust for a new management theory. *Asian J. Manag.* 2017, 8, 921–924. [CrossRef]

32. Andrienko, G.; Andrienko, N.; Rinzivillo, S.; Nanni, M.; Pedreschi, D.; Giannotti, F. Interactive Visual Clustering of Large Collections of trajectories. In Proceedings of the 2009 IEEE Symposium on Visual Analytics Science and Technology, Atlantic City, NJ, USA, 11–16 October 2009; IEEE: Manhattan, NY, USA, 2009; pp. 3–10.

33. Wu, X.; Zurita-Milla, R.; Kraak, M.-J. Co-clustering geo-referenced time series: Exploring spatio-temporal patterns in Dutch temperature data. *Int. J. Geogr. Inf. Sci.* 2015, 29, 624–642. [CrossRef]
34. Han, J.; Kamber, M.; Pei, J. Data mining: Concepts and techniques, Waltham, MA. *Morgan Kaufman Publ.* 2012, 10, 971–978.
35. Banerjee, A.; Dhillon, I.; Ghosh, J.; Merugu, S.; Modha, D.S. A generalized maximum entropy approach to Bregman Co-clustering and matrix approximation. *J. Mach. Learn. Res.* 2007, 8, 1919–1986.
36. Cho, H.; Dhillon, I.S.; Guan, Y.; Sra, S. Minimum sum-squared residue co-clustering of gene expression data. In Proceedings of the 2004 SIAM International Conference on Data Mining, 22–24 April 2004; SIAM: Lake Buena Vista, FL, USA, 2004; pp. 114–125.
37. Wu, X.; Zurita-Milla, R.; Kraak, M.J. A novel analysis of spring phenological patterns over Europe based on co-clustering. *J. Geophys. Res. Biogeosci.* 2016, 121, 1434–1448. [CrossRef]
38. Head, K.; Mayer, T.; Ries, J. The erosion of colonial trade linkages after independence. *J. Int. Econ.* 2010, 81, 1–14. [CrossRef]
39. Hou, J. Revisiting the trade effects of the euro: Data sources and various samples. *Empir. Econ.* 2020, 59, 2731–2777. [CrossRef]