Trade without "proportionality": A novel approach to trace import uses

Yongbiao Fang, Zhaoyang Jin, and Jiansuo Pei*
(School of Applied Economics, Renmin University of China)

Abstract: Import share is a key parameter in analysis of international trade. However, previous research mainly relies on the "proportionality" assumption (and its variants) to obtain share parameters, which may lead to biased estimates. In contrast, this paper introduces a novel approach that combines machine learning techniques with big data to accurately trace import uses within narrowly defined industries. We present a novel methodology to estimate the import matrix without "proportionality" assumption. To this end, based on Chinese customs product trade data, enterprise business registration data and rich micro-survey data, this study develops machine learning algorithms to precisely trace import uses. This approach is then applied to compile China's non-competitive input-output tables with 37 sectors during the period of 2000-2016. Essentially, our novel estimation approach overcomes the limitations of the "proportionality" assumption by utilizing rich micro datasets to identify the source sector and the use sector of each product. The procedure is as follows. Using customs trade data at product level, we first match it according to the concordance table between HS 8-digit code and the input-output sector to identify the source sector of the product. Then, we match it with the enterprise business registration database to identify the sector of the enterprise, which represents the use sector of the product. In this way, our estimation method directly traces the end-uses of imports rather than imposing "proportionality" assumption. Also, we compare the newly estimated import matrix with counterparts relying on the "proportionality" assumption. The comparison results show that, overall, the differences in most import share estimates between the two methods are not significant. However, there are also some sectors with large differences in share estimates, exceeding 50 percentage points. We conclude by generalizing this methodology, which can be applied and adapted to other economies when tracing import uses, and serving as key inputs for analysis of international trade.

Keywords: import share, machine learning, proportionality assumption, intermediate imports

JEL Codes: F12; F14; C67; C81

1. Introduction

Since the 1980s, with the upgrade of transportation technology and the continuous reduction of trade costs, Global Value Chains (GVCs), which are mainly characterized by production fragmentation and trade integration, have become the dominant organizational form of global production. The production of goods and services is vertically split into multiple processes, arranged in different countries around the world according to comparative advantages, and assembled sequentially along the supply chain or at the final destination. Production fragmentation is accompanied by the scale expansion and structure transformation of international trade, especially with intermediate goods trade occupying an increasingly important position. According to data from the WTO, the trade volume of intermediate goods accounted for more than 50% of the total global trade (excluding energy) in the decade from 2011 to 2021. More and more companies began to use imported components in their production processes, while an increasing number of component manufacturers began to export their products to the international market. GVCs have brought global economies closer together, profoundly impacting the global economic development, the competitive pattern between countries, and the business model of companies. At the same time, they also pose new challenges to traditional theories and accounting methods of international trade.

A vast body of literature has studied the new type of international trade under GVCs and its impacts. Including but not limited to, the measurement of trade in value added and the degree of GVCs participation (Feenstra and Hanson, 1996; Hummels et al., 2001; Johnson and Noguera, 2012; Koopman et al., 2014), the measurement of position within GVCs (Antràs et al., 2012), the theoretical modeling of international trade under GVCs (Caliendo and Parro, 2015; Alfaro et al., 2019; Antràs and de Gortari, 2020), and the assessment of the economic impacts of GVCs based on various measurement indicators, such as the impact of specialization in GVCs on the labor market, total factor productivity, and technological spillovers (Feenstra and Hanson, 1999; Ertur and Koch, 2007; Houseman et al., 2011; Timmer et al., 2013). The above studies effectively supplement the research on international trade under the framework of global division of labor, but all of them are limited by data quality to some extent. The sectoral trade share is a core indicator for these studies, and the accuracy and continuity of this data directly determine the reliability of relevant studies. Most of the existing literature obtains data on sectoral trade shares from national input-output tables and the World Input-Output Database (WIOD), OECD Inter-Country Input-Output (OECD ICIO) Tables, and other ICIO sources. However, most countries, including China and the United States, do not track whether imports are for final demand or intermediate use, nor do they investigate the information of each sector's imports of each input. Therefore, researchers have used two important proportionality assumptions in their studies. One is that imports per sector are split into final and intermediate good use in the same proportion as is the case on the national level of imports. Secondly, intermediate imports are split across purchasing sectors in proportion to their overall

imported intermediate use, i.e. the import proportionality assumption.[1] However, the "proportionality assumption" is an overly simplified hypothesis that hardly reflects the reality accurately and greatly affects the reliability of relevant research.

In terms of accuracy, the National Research Council (2006) critiqued the proportionality assumption as a significant limitation of current data collection and analysis. The research of Chen et al. (2020) on China and the research of Timmer et al. (2015) on the world both found noticeable differences in the trade in value added and other indicators calculated by using the input-output tables compiled based on various assumptions, including those from WIOD. Dean et al. (2011) distinguished between intermediate and final products in China's imports by introducing the United Nations Broad Economic Categories (BEC) classification and detailed product-level trade flows. This approach has been widely used in similar works by WIOD (Dietzenbacher et al., 2013), the OECD-ICIO (Koopman et al., 2014), as well as the most recent GTAP editions (Carrico et al., 2020), largely addressing the issues caused by the first proportionality assumption. However, the import proportionality assumption is still widely used in estimating a sector's imports for each type of input. To address this issue, researchers have begun to use more detailed micro data to directly estimate the sectoral import matrices. In addition, existing studies on the United States (Feenstra and Jensen, 2012), Germany (Winkler and Milberg, 2012), and Asia (Oosterhaven et al., 2008; Puzzello, 2012) have all found that the import proportionality assumption does not hold when compared with the reality. Starting with micro data, these studies provide valuable insights into alternatives to the import proportionality assumption. However, since all the three studies are based on special data, they are practically infeasible to apply and generalize. Moreover, most of the data in these studies are sampling data, and under equal conditions, the higher the proportion of sampled enterprises' imports in total imports, the higher the accuracy of the estimate. Feenstra and Jensen (2012) reported this proportion to be 75% and suggested that such a proportion might reduce the accuracy of import matrix estimates. The IDE-JETRO's Asian International Input-Output (AIIO) Tables compiled by Puzzello (2012), based on enterprise survey method, are almost impossible to achieve this proportion in sample surveys.

Additionally, in terms of continuity, although databases such as WIOD, OECD ICIO, and Eora MRIO provide continuous input-output data that can facilitate continuous research, their compilation methods primarily rely on the non-continuous national input-output tables (most of which are non-competitive tables) published by countries around the world. They first integrate macroeconomic statistical data and use mathematical methods to estimate and supplement for the intervening years, then combine trade data and the proportionality assumption to generate the final input-output tables. While addressing the issue of discontinuity in official data, this approach may further reduce the accuracy of the data.

In response to the above situation, this paper proposes a method to estimate the sectoral import matrices based on corporate registration data and trade data, which addresses the dependence on the import proportionality assumption of existing data

[1] In the text that follows, "proportionality assumption" specifically refers to the first assumption, and "import proportionality assumption" specifically refers to the second assumption.

while considering the continuity of the data. It also applies machine learning methods to improve the credibility of the estimates. Based on this method, this paper also compiles China's time-series non-competitive input-output tables during the period of 2000-2016. Compared to existing research, this paper makes the following marginal contributions: First, the corporate registration data and customs trade data used in this paper are continuous and comprehensive, which are also statistically gathered by most countries worldwide, offering stronger generalizability. This improves the accuracy and timeliness of the import matrix estimates and facilitates widespread adoption. Second, this paper employs machine learning methods to identify the sectors to which companies belong based on their business scope, effectively increasing the proportion of samples available for estimation. Additionally, to facilitate practice in different countries, this paper constructs mature programs and training sets. Third, this paper effectively supplements the relevant scenarios for China, the largest trading nation and the largest developing country. In summary, by combining micro data and machine learning methods, this paper proposes a more accurate and easily generalizable method for estimating import matrices, which will greatly benefit research on international trade under GVCs.

2. Basic framework and technology

The main content of this section is to summarize the basic framework and techniques of estimating intermediate products import share matrices (IISMs) and compiling non-competitive input-output tables, and introduce the sector classification method and table format used in this paper.

In the process of estimating IISMs, we adopt different estimation methods for the goods sectors (the first 22 sectors) and the service sectors (the last 15 sectors). The estimation of IISMs of the goods sectors is based on China’s customs trade data and corporate registration data, combined with deep learning algorithms. Using customs trade data at product level, we first match it according to the concordance table between HS 8-digit codes and the input-output sectors to identify the source sector of the imported products. Then, we match it with the enterprise business registration database to identify the sector of the enterprise, thus obtaining the distribution of imported products among various sectors, i.e., the sectoral trade shares. The estimation of import shares of the services sectors is obtained directly based on the import proportionality assumption. Using this estimation method, the estimated import coefficient matrices for both goods and services sectors can be obtained.

When identifying the use of imported products, to improve the previous method of splitting the use of imported products according to the proportionality assumption, we adopt a method similar to Timmer et al. (2013). This involves directly determining the use of goods according to the BEC classification. Based on the customs trade data, the proportions of each import used for intermediate use, consumption, and fixed capital formation are estimated according to the BEC classification. Then combined with sector-specific import value data calculated by Zhang et al. (2021), the total amounts of each import used for intermediate use, consumption, and fixed capital formation can be obtained. By further utilizing the estimated IISMs, the amount of imported

intermediate products from each sector used for each sector can be calculated, which are, the intermediate products import flow matrices (IIFMs). Embedding them into the competitive input-output tables, with some necessary adjustments, we can obtain non-competitive input-output tables.

Considering the available data, this paper chooses to estimate the IIFM for 2000-2016 and compile non-competitive input-output tables for the same period, using the China's time-series competitive input-output tables compiled by Zhang et al. (2021). Considering that the widely used method for estimating import matrices is based on the import proportionality assumption to split the imports into different uses, whereas this paper primarily estimates based on micro-databases. Therefore, this paper also compares the matrices obtained using these two different estimation methods. The basic research idea and framework are illustrated in Figure 1.

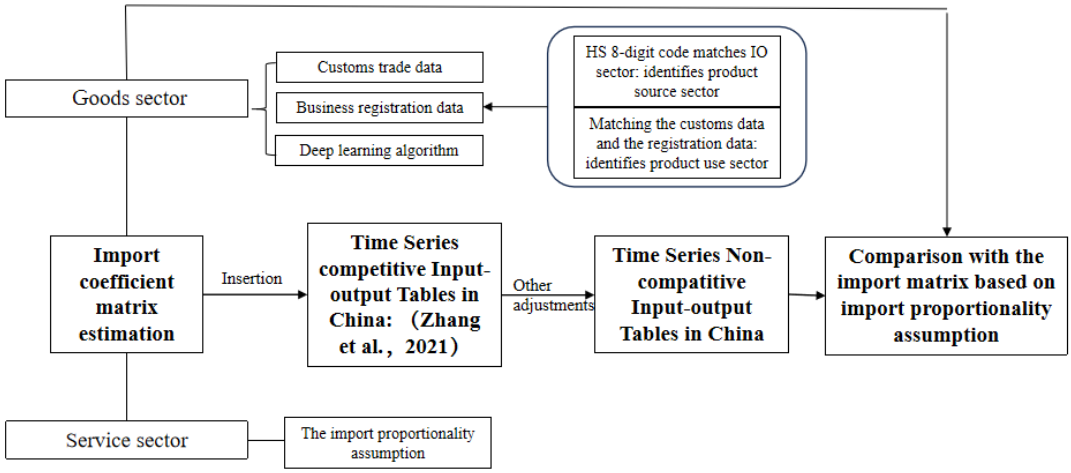


Figure 1 The basic research framework of this paper

The method of sector classification in input-output tables is not unique, so it is necessary to determine the classification of input-output sectors before compilation. The National Bureau of Statistics is responsible for compiling China's benchmark year competitive input-output tables. However, since their compilation relies on specialized input-output surveys, which require a significant amount of human, material, and time resources, the intervals between benchmark year input-output tables are relatively long, with compilation years falling on those ending in 2 and 7. In addition to the benchmark year input-output tables, the National Bureau of Statistics also compiles extended input-output tables in years ending in 0 and 5, based on the benchmark survey year tables. Due to revisions in China's national economic industry classification in 2002 and 2011, the sector classification in the input-output tables published by the National Bureau of Statistics was also revised in relevant compilation years, but the overall number of input-output sectors has always been maintained at over 100. Moreover, the WIOD input-output tables include 42 sectors, while the OECD input-output tables include 45 sectors. Zhang et al. (2021) compiled a series of input-output tables for China from 1981 to 2018, adopting a unified sector classification with a total of 37 sectors. The principle of their sector classification is: first, referring to the sector classification of input-output tables and extended tables published by the National Bureau of Statistics at different stages to determine the corresponding sector classification; then,

determining a unified sector classification for the entire series during consistency adjustments across the whole series. For the estimation of the IISMs in this paper, the correspondence between HS 8-digit codes from customs and sectors from the National Bureau of Statistics' benchmark year input-output tables was used, as well as the industry classification of companies in the corporate registration database. Therefore, the determination of input-output sectors in this paper needs to consider both the sector classification of the benchmark year input-output tables and the industry classification of the corporate registration database, ensuring that the final sector classification corresponds to both classifications. After estimating the IISMs, further calculation of the import matrix will be conducted, and it will be embedded into the competitive input-output tables series compiled by Zhang et al. (2021) to obtain a series of non-competitive input-output tables. Thus, considering the above factors, the final sector classification of the input-output tables in this paper follows the method of Zhang et al. (2021), with specific sector classifications presented in the appendix.

After clarifying the sector classification of the input-output tables, the format of the input-output tables can be determined. Compared to competitive input-output tables, the difference in non-competitive input-output tables lies in distinguishing between domestic products and imported goods in both intermediate and final uses, thereby determining the format of non-competitive input-output tables as shown in Table 1. The most important part of the non-competitive input-output table is the import matrix $\boldsymbol{Z}$, which is a crucial part of the input-output table depicting the distribution and use of imported goods across sectors. Its accuracy directly determines the accuracy of analyses of China's international trade (measurement of GVCs). Let *i* and *j* respectively represent the source sectors of intermediate inputs and the use sectors of intermediate inputs, then let Z_{ij} represents the portion of imports from sector *i* used for intermediate consumption in sector *j*, and $Z_i = \sum_j Z_{ij}$ represents the total amount of imports from sector *i* used for intermediate consumption. Letting $\omega_{ij} = \frac{Z_{ij}}{Z_i}$ represent the proportion of imports from sector *i* used for intermediate consumption in sector *j*, then $\Omega=(\omega_{ij})$ is the IISMs. Furthermore, C_i and K_i respectively represent the portions of imports from sector *i* used for consumption and fixed capital formation. Letting $M_i = Z_i+C_i+K_i$ represents the total amount of imports from sector i, then $z_i = \frac{Z_i}{M_i}, c_i = \frac{C_i}{M_i}$ and $k_i = \frac{K_j}{M_i}$ respectively represent the proportions of imports from sector *i* used for intermediate use, consumption, and fixed capital formation. Subsequent text uses corresponding letters with a prime to denote estimated quantities.

Table 1 The format of non-competitive input-output table

Output / Input				Intermediate use					Final use				Rests	Domestic output/Import
		Sector name		Sector 1	Sector 2	…	Sector 37	Total intermediate use	Consumption	Fixed capital formation	Inventory change	Export		
Intermediate input	Domestic product	Sector N												
	Import product	Goods sector	Sector 1	Z_{11}	Z_{12}	…	$Z_{1,37}$	Z_1	C_1	K_1				
			Sector 2	Z_{21}	Z_{22}	…	$Z_{2,37}$	Z_2	C_2	K_2				
			…	…	…	…	…	…	…	…				
			Sector 22	$Z_{22,1}$	$Z_{22,2}$		$Z_{22,37}$	Z_{22}	C_{22}	K_{22}				
		Service sector	Sector 23	$Z_{23,1}$	$Z_{23,2}$		$Z_{23,37}$	Z_{23}	C_{23}	K_{23}				
			…	…	…	…	…	…	…	…				
			Sector 37	$Z_{37,1}$	$Z_{37,21}$	…	$Z_{37,37}$	Z_{37}	C_{37}	K_{37}				
	Total intermediate input													
Initial input	Factor N													
Total input														

3. Data

3.1 Primary data sources

The databases primarily used in this paper are the China's Customs Trade Database (CTD) from 2000 to 2016, and the 2022 version of the China's Industrial and Commercial Registration Enterprise Database (CICD) and Time Series Input-Output Tables in China (Zhang et al., 2021).

3.1.1 China's Customs Trade Database from 2000 to 2016

The China's Customs Trade Database is an authoritative data source released by the General Administration of Customs of China. It contains detailed records of every import and export product traded by enterprises, with each product identified by an HS 8-digit code. For each product, the data includes three types of information: the first type provides basic trade variables such as trade amount, trade status (import/export), and number of products traded; the second type provides variables related to trade patterns and methods, including the country or region of import or export, trade mode, transportation mode, and transshipment country; the third type provides basic information about trading companies, including their names, customs codes, and, in the case of data from 2000 to 2006, their addresses, telephone numbers, and email addresses. Due to the database's authority and the richness of the indicators it contains, it has been widely used in empirical research on various issues related to China's imports and exports (Yu, 2015; Kee and Tang, 2016). The time span of the customs trade data used in this paper is from 2000 to 2016, and only includes import data records. The variables used include import or export, enterprise name, commodity HS code, trade amount, trade mode, telephone, and email.

3.1.2 China's Industrial and Commercial Registered Enterprises Database (2022 version)

The China Industrial and Commercial Registration Enterprise Database is managed and compiled by the State Administration for Industry and Commerce, and records basic registration information for enterprises. The variables included are: enterprise name, former name, unified social credit code, registered address, industry, business scope, business status, registration authority, establishment time, telephone, email, etc. Enterprises are one of the most important microeconomic units in modern economic activities, so micro-level data on enterprises is a core material required for research in various fields of microeconomics. As a database theoretically containing basic information for all registered enterprises, the China Industrial and Commercial Registration Enterprise Database can also act as a bridge connecting with other enterprise databases, and the integration of different databases can provide researchers with more research perspectives. The use of this database is also an innovation point of this paper, as there are still relatively few studies using it to investigate trade-related issues. The industrial and commercial registration database used in this paper is the 2022 version, which includes approximately 178 million entries of registered industrial and commercial enterprises from all provinces, municipalities, autonomous regions, and special administrative regions in China from 1949 to 2022. The variables used include enterprise name, former name, industry, business scope, telephone, and email.

Since the database only provides one former name and one current name for each enterprise, but some enterprises may have more than one former name, changes exist naturally between different versions of the database. Therefore, when using the database, the more recent the version is to the year needed, the more accurate information it provides.

3.1.3 Time Series Input-Output Tables in China (Zhang et al., 2021)

Zhang et al. (2021) compiled the time series competitive input-output tables in China from 1981 to 2018, and calculated the total import of sub-sectors during the compilation process. In this paper, we use the import volume data of the sub-sector products when calculating the intermediate flow matrix of imports, and use the complete competitive input-output tables when compiling the non-competitive input-output tables.

3.2 Initial data processing

3.2.1 Processing of the customs trade database

(1) Exclude data with missing key variables

Since this paper focuses on issues related to imports, only import data from the customs trade database is retained. The customs trade database contains a wealth of information, but due to various reasons, some key variables have missing values. In this study, it is necessary to match the enterprise names and HS codes in the customs trade database with other data, and calculate the trade volume. Therefore, to proceed smoothly in the next steps, this study excludes observations with missing values for enterprise name, commodity HS code, and transaction amount. The proportion of import records with missing values in the customs database from 2000 to 2016 is shown in the first column of Table 2.

(2) Exclude data with trade mode of processing and assembly trade of supplied materials

In the customs trade database, there is a trade mode called processing and assembly trade of supplied materials. This trade mode is a type of processing trade, where one country provides raw materials or components, which are then processed and assembled by another country, and the finished products are exported to a third country or returned to the country of origin of the raw materials. This trade mode only involves the trade of processing and assembly services, without involving actual goods. Therefore, when estimating the IISMs, trade records with the trade mode of processing and assembly trade of supplied materials need to be excluded. The classification of trade modes in the customs trade data from 2000 to 2006 is more detailed, explicitly distinguishing processing trade into processing trade with supplied materials and processing trade with supplied inputs. Therefore, trade records with the trade mode of processing and assembly trade of supplied materials can be directly excluded.

After the data processing mentioned above, the basic situation of the customs trade database obtained is shown in Table 2. Among them, columns (2) and (3) respectively represent the number of enterprises that have imported goods and the total import value of products (unit: USD 100 million) after the above processing.

Table 2 Basic information of customs trade database

Year	(1) Share of import records with missing values	(2) Number of goods importing enterprises	(3) Total product imports (100 million dollars)
2000	0.60%	59505	1947.99
2001	0.63%	63587	2303.67
2002	1.14%	70003	2445.77
2003	0.00%	78734	3511.46
2004	2.63%	88749	5037.93
2005	14.66%	84198	5403.45
2006	1.09%	113100	7163.81
2007	7.69%	119197	8877.08
2008	5.71%	126489	10720.36
2009	3.34%	130109	9746.46
2010	1.70%	142256	13586.03
2011	0.59%	153094	16909.90
2012	0.12%	159029	17793.43
2013	0.13%	164298	19009.41
2014	0.45%	164985	19050.66
2015	6.16%	148167	15624.25
2016	0.90%	270582	18606.17

3.2.2 Processing of industrial and commercial registered enterprises database

The China's industrial and commercial enterprise registration database contains records with missing company names and duplicate records. In order to improve the accuracy and efficiency of data processing, these records have been removed. After the removal, the proportion of remaining enterprise registration information records to the original records is approximately 99%. To facilitate data processing, only a few variables needed from the industrial and commercial enterprise registration database are retained: company name, former name, industry, business scope, telephone, and email.

This database is primarily used to identify the industry to which an enterprise belongs and its input-output sector. The business scope of the enterprise is additional auxiliary identifying information. The industry classification used in this paper's industrial and commercial enterprise registration database consists of a total of 127 categories, which can be manually corresponded to 37 input-output sectors.[1] However, the industry classification standards in this database are not uniform, and some industry classifications are too broad, resulting in cases where one industry corresponds to multiple input-output sectors. For example, the mining industry corresponds to both "metal ore mining and dressing products" and "non-metallic and other mineral mining and dressing products," and the manufacturing industry corresponds to even more

[1] See the appendix for the table of correspondence between industrial and commercial industries and 37 input-output sectors.

sectors. Enterprises with unclear sectoral affiliations account for approximately 1.04% of the entire industrial and commercial enterprise registration database. Additionally, there are records with missing industry information in the database, accounting for approximately 14.5% of the total.

In order to increase the sample size and improve the accuracy of estimation as much as possible, further processing is required for enterprises whose input-output sectors cannot be identified. Machine learning is an important branch in the field of artificial intelligence, and its core idea is to construct models based on training data and then use these models for prediction or decision-making. Classical machine learning is typically divided into two categories: supervised learning and unsupervised learning. Supervised learning is further divided into two types: classification and regression. The process of using machine learning for prediction or classification involves using labeled training set data to identify and extract features, establish prediction models, and apply the learned patterns to new data for prediction or classification. In the digital world, most of the data is unstructured, especially text data, which is in need of NLP as a "bridge" between human and machine for communication. Natural Language Processing (NLP) is an interdisciplinary field combining computer science, artificial intelligence, and linguistics. Its core task is to convert human language into a form that computers can understand and process, enabling the transformation of large and complex text data into analyzable and usable information. With the development and application of machine learning in the field of natural language processing, its efficient and accurate predictive performance plays a crucial role in assisting economic measurement, especially in the face of missing important data.

It is noted that besides the industry classification information of enterprises, the industrial and commercial registration database also includes the variable of business scope, and there exists a corresponding relationship between the business scope and industry or sector classification of the enterprises. The amount of data in the industrial and commercial registration database is large, and the existing correspondences between the business scope and the sectors to which the enterprises belong are relatively accurate. Therefore, machine learning algorithms can be utilized to construct a model using the existing correspondences between the business scope and input-output sectors as training data. Subsequently, based on this constructed model, the sectors of enterprises with known business scope but unknown sector can be predicted.

For better understanding, a simplified explanation of the classification task is as follows: Given a text data set $D = \{T_1, T_2, \dots, T_n\}$ and the corresponding categories $Y = \{y_1, y_2, \dots, y_k\}$, the aim is to train a model M to predict the value $Y_i = M(T_i)$ for a given data set that is as close to the true value as possible. As this paper deals with a total of 37 categories of data, the category $Y \in [1,\ 2, \dots,\ 37]$. Only the descriptions of enterprises' business scope and their corresponding sector categories are retained during the machine learning process, with sector category codes consistent with the sector classification codes in the appendix. In order to make full use of data resources and mitigate the risk of overfitting at the same time, this paper directly uses the text representations generated by the pre-training model as input during the experimental process, and then fine-tuning through the sequential model, and finally uses the fully

connected neural network for classification. The whole process mainly consists of five stages: data preprocessing, model design, model training, model evaluation, and missing data completion.

When evaluating the performance of classification models, *Precision* is a commonly used academic metric to measure the accuracy of the model when predicting positive categories, defined as follows:

$$Precision = TP / (TP + FP)$$

Here, *TP* (True Positives) represents the number of samples correctly predicted as positive categories by the model, and *FP* (False Positives) represents the number of negative samples incorrectly predicted as positive categories by the model. Precision measures the accuracy of the model in predicting positive categories, i.e., how many samples are actually true positive categories when the model predicts positive categories, and it is widely used in information retrieval, machine learning, statistics, and other fields. This paper first uses this indicator to measure the prediction effect of each sector category, and then uses the arithmetic mean (macro-precision) to make a reasonable evaluation of the overall effect of the model, as shown in specific results in Table 3. [1]

Table 3 The result of the precision

Indicator	Category								
Precision	01	02	03	04	05	06	07	08	09
	0.97	0.96	0.75	0.91	0.97	0.96	0.94	0.98	0.96
	10	11	12	13	14	15	16	17	18
	0.90	0.96	0.85	0.88	0.87	0.82	0.81	0.86	0.93
	19	20	21	22	23	24	25	26	27
	0.91	0.86	0.86	0.93	0.94	0.94	0.98	0.95	0.94
	28	29	30	31	32	33	34	35	36
	0.79	0.95	0.97	0.87	0.97	0.95	0.78	0.94	0.86
	37		**Macro-precision**						
	0.77		**0.904**						

The results show that the macro-precision of the model is 0.904, indicating that the probability of correctly predicting true positive cases (macro-precision) is over 90% for each predicted category. As shown in the appendix, the macro-recall is 0.89 and the macro F1-score is 0.89, indicating that the model can effectively capture positive samples (macro-recall) and demonstrate excellent predictive performance in terms of accuracy and completeness (macro F1-score). Since the model trained on the existing dataset is relatively reliable, this paper freezes the model parameters to predict and complete the missing sector data.[2]

For the part of the industrial and commercial registration database that already has detailed industry classification information, it can be matched with the industry-to-

[1] In addition, Recall and F1-score are also commonly used metrics. Recall measures the model's ability to find all truly positive samples, while F1-score is the harmonic mean of Precision and Recall, aiming to comprehensively consider the accuracy and completeness of the model. The sectoral results and macro results of Recall and F1-score are shown in the appendix.

[2] Please refer to the appendix for the detailed process of predicting enterprise sectors based on business scope using machine learning algorithms.

input-output 37-sector reference table manually to determine the sector to which the enterprise belongs. For the part where the industry classification information is missing or too broad, a model has been constructed using the machine learning method mentioned above based on the existing data to predict their sectors. Combining these two parts together yields the correspondence between enterprises in the industrial and commercial registration database and the input-output 37 sectors. Finally, after the above processing, the proportion of enterprises in the industrial and commercial registration database that can be identified with their corresponding input-output sectors is approximately 99.38%, which has been greatly improved compared with before the processing.

4. Imported intermediate products, consumption, and fixed capital formation

In the competitive input-output tables, the products used are not distinguished from domestic products or imported products. Imports are shown as a separate column, indicating the total amount of imports for each product sector, without specific differentiation of the destination of imported products. In the non-competitive input-output tables, domestic products and imported products are detailed separately. In this case, the import section in the table changes from a single column to a matrix, providing a detailed breakdown of the destinations for various imported products. When dividing imported goods into different uses, it is assumed that imported goods are only used for intermediate use, consumption, and fixed capital formation, excluding considerations for imported goods in inventory changes and re-exports. This section primarily focuses on estimating the proportions of imports from each sector used for intermediate use, consumption and fixed capital formation (z_i, c_i and k_i), and then further obtain the total amounts of imports from each sector used for these three directions (Z_i, C_i and K_i) based on the total import volume of products by sector, so as to prepare for the next step of splitting the original competitive input-output tables into non-competitive input-output tables.

The main data basis for this section are the customs trade data after the previous initial data processing and the sector-specific import value data from Zhang et al. (2021).[1] The basic idea is to split the total imports by sector according to their use, with the import value data sourced from Zhang et al. (2021). For the estimation of the proportions of various usage destinations of imported products, the goods sectors primarily rely on customs trade data. The source sectors of imported goods are identified based on the concordance table between customs HS 8-digit codes and input-output sectors. And the usage destinations of imported products (intermediate use, consumption, or fixed capital formation) are identified based on the concordance table between customs HS 6-digit codes and the BEC commodity classification standards. As for service sectors, the proportions of imported products directed to each usage

[1] The calculation method of total import in Zhang et al. (2021) is "total import = total import of customs goods + goods and services directly purchased by Chinese residents abroad + services provided by foreign residents to Chinese residents," which is more comprehensive than the total import of direct customs goods.

destination are directly estimated according to the proportionality assumption.

For the identification of the usage destinations of imported goods , we refer to the method proposed by Timmer et al. (2013) to split the usage destinations based on the BEC product classification standard.[1] First of all, for the customs trade data from 2000 to 2006 with a more detailed classification of trade modes, the products classified under the processing trade of imported materials are identified as being used for intermediate use, and the products classified under processing trade of imported equipment are identified as being used for fixed capital formation. Then, for the identification of usage destinations of other imported products, the usage destination of imports is divided into three categories according to their BEC categories: intermediate use, consumption and fixed capital formation. The customs trade database contains the information of HS 8-digit code of the product, and truncating the first 6 digits yields the product's HS 6-digit code. Since the HS codes underwent three changes in 2002, 2007, and 2012, only the benchmark years (1996, 2002, 2007, and 2012) versions of the HS codes are available, and the HS codes used in non-benchmark years are the latest versions of the HS codes available in those years.[2]

The United Nations Statistics Division (UNSD) provides the comparison table of customs HS 6-digit codes and BEC codes for benchmark years (1996, 2002, 2007, and 2012). By matching customs trade data with the corresponding benchmark year's comparison table, the BEC code of imported products can be obtained, thereby identifying the usage destination of the products (intermediate use, consumption, or fixed capital formation). Additionally, the comparison table may contain cases where a single product corresponds to two or more BEC codes, in which the use of the product is divided into two or more categories. For example, in most cases, starch and wheat are used as intermediates for producing other products, but in certain instances, they may be directly consumed as final products. The comparison table includes the proportions of the various uses of the products. For products that are simultaneously allocated to multiple usage categories, the import value of such products in the customs trade database needs to be split according to their respective proportions and aggregated into the corresponding usage categories.[3]

For the identification of the source sectors of imported products in the customs trade database, it is primarily based on the comparison table between the HS 8-digit code of the product and the input-output sectors in the input-output table published by the NBS in the base year, and the comparison table between the input-output sectors in the base year and the 37 input-output sectors in this paper. By matching the customs trade data with the comparison table between the HS 8-digit code and the input-output sector, the source sector of the imported products can be obtained. In the comparison table between HS 8-digit code and input-output sectors, the input-output sectors are

---

[1] See the UNSD Fifth Edition (BEC 5) document revised in 2016 for details. The BEC Product Classification standard is a set of internationally accepted commodity classification standards developed by the United Nations Statistical Division (UNSD), which classifies commodities into different categories according to their characteristics, functions and uses.

[2] In other words, the 1996 version of the HS8 code was used for the years 2000-2001, the 2002 version for the years 2002-2006, the 2007 version for the years 2007-2011, and the 2012 version for the years 2012-2016.

[3] The comparison table provides the split ratios of different uses of products, which, although it may be different from the reality in China, is a relatively effective method of handling aside from specific input-output surveys.

over 100 sectors from the competitive input-output tables published by the NBS for the benchmark year, which then need to be further matched with the comparison table between the input-output sectors in the base year and the 37 input-output sectors in this paper to obtain the required 37 input-output sectors.

It should be noted that in the comparison table between HS 8-digit code and input-output sectors, there are cases where one HS 8-digit code corresponds to two input-output sectors. Furthermore, these two corresponding input-output sectors may belong to the same 37 input-output sectors, or belong to two different 37 input-output sectors. For the latter case, the import records for the products with these HS 8-digit codes in the customs trade database need to be handled separately: they should be split according to the ratio of the total intermediate use of the corresponding input-output sector in the competitive input-output table for the respective year. Specific split ratios can be found in the attached table. By performing the aforementioned splits and matches for customs trade import data from 2000 to 2016, the importing sectors of the products can be determined, i.e., the products' source sectors. The identification rates for the product source sectors in the customs trade database for each year are shown in Table 4 below.

Table 4 Identification rate of source sector for imported goods in customs trade database

Year	(1) The number of import records	(2) Import value of goods
2002	99.57%	99.61%
2003	96.84%	88.06%
2004	96.39%	86.70%
2005	94.88%	83.53%
2006	93.59%	79.04%
2007	100.00%	100.00%
2008	97.48%	97.21%
2009	95.07%	93.23%
2010	93.87%	92.30%
2011	92.74%	91.41%
2012	99.99%	98.33%
2013	99.53%	96.86%
2014	97.32%	96.62%
2015	97.15%	96.45%
2016	93.85%	97.35%

For the estimation of the proportions of imports for intermediate use, consumption, and fixed capital formation across various product sectors, different estimation methods are employed for goods and service sectors due to the nature of trade records contained in the customs trade database, which primarily include physical commodity trade records. For the goods sectors, by summarizing the matched customs data according to the usage destination and the source sector, we can get Z_i', C_i' and K_i', with $M_i' = Z_i' + C_i' + K_i'$. According to $z_i = \frac{Z_i'}{M_i'}$, $c_i = \frac{C_i'}{M_i'}$ and $k_i = \frac{K_i'}{M_i'}$, the proportions of intermediate

use, consumption, and fixed capital formation for each goods sector can be obtained.[1] For the service sector, a proportional assumption is made for the splitting of imports across different usage categories, directly providing the proportions for intermediate use, consumption, and fixed capital formation for each imported product. Finally, we can directly obtain the the proportions of each imported product used for intermediate use, consumption and fixed capital formation by splitting the imported products according to the assumption of equal proportions in different uses. Finally, since it is assumed that all imported goods are used for the three purposes of intermediate use, consumption and fixed capital formation, based on the import value data of different sectors in Zhang et al. (2021), namely M_i, the total amount of imported goods used for these three purposes($\mathrm{Z_i}'$, $\mathrm{C_i}'$ and $\mathrm{K_i}'$) can be obtained according to $Z_i = M_i z_i$, $C_i = M_i c_i$ and $K_i = M_i k_i$.

5. Estimation of IISMs

This paper mainly relies on the China's customs trade database for the estimation of the IISMs, which mainly contains import and export trade data for physical products and lacks the data records for service trade. The product source industries corresponding to imported physical goods are agriculture and industry, while the product source industry for imported services is the service sectors. Therefore, the input-output sectors correspondingly divide into the goods sectors and the service sectors. As a result, different estimation methods need to be adopted for the IISMs of the goods sectors and the service sectors: The estimation of the IISMs for the goods sectors is based on the China's Customs Trade Database and the Industrial and Commercial Enterprise Registration Database, while the estimation of the IISMs for the service sectors depends on the import proportionality assumption.

The estimation of the IISMs in the goods sectors is mainly based on the customs trade database. During the estimation process, only the import of intermediate products is involved, so it is necessary to retain only the data records of imported goods classified as intermediate products in the customs trade database. In the previous section 4, imported goods were categorized into three groups according to their use: for intermediate use, consumption, and capital formation. Goods used for intermediate use are considered intermediate products, so in this section of the estimation, only the trade records with goods classified as intermediate use need to be retained from the customs trade database. In addition, in the customs trade database, there is a portion of enterprises that are specifically engaged in commodity trade. These enterprises do not have their own production activities but act as intermediaries: they import goods from abroad and then sell them to other domestic enterprises. The intermediate products imported by these enterprises are mostly not used for their own production activities but first enter other enterprises as goods and then participate in the production activities of those enterprises. An important assumption underlying the estimation of the IISMs in the goods sectors is that the intermediate products imported by enterprises are used for their own production activities. Therefore, in order to improve the accuracy of the estimation, it is necessary to exclude such trading enterprises from the sample.

[1] The corresponding letter plus a prime is used to represent the estimate based on the customs trade database.

Following the approach used by Ahn et al. (2011) to identify trading enterprises, data records with enterprise names containing "import and export," "trade," "economic trade," "foreign economic," and "trade" in the customs trade database are ultimately removed.

5.1 Goods sectors

The estimation of the IISMs for the goods sectors is based on the initially processed customs trade database. First, according to the concordance table between HS 8-digit codes and the input-output sectors, we can obtain the source sectors of each product recorded in the initially processed customs trade database. Then, by matching the initially processed customs trade data with the initially processed industrial and commercial enterprise data, the sectors to which the importing companies belong can be obtained. It is assumed that the intermediate products imported by companies are used as intermediate inputs for their own production, so the sector to which the company belongs is considered as the use sector for its imported intermediate products. Finally, the trade volume in the initially processed customs trade database is grouped and summed according to the source sectors and use sectors of the products, yielding the estimate Z_{ij}'. According to the equations $Z_i' = \sum_j Z_{ij}'$ and $\omega_{ij}' = \frac{Z_{ij}'}{Z_i'}$, we can obtain the sectoral import shares of the goods sectors.

5.1.1 Identification of the source sector

For the identification of the source sectors of the goods in the customs trade database, the same method as that used in section 4 for identifying source sectors is employed. This primarily involves the comparison table between the HS 8-digit code of the product and the input-output sectors in the input-output table published by the NBS in the base year, and the comparison table between the input-output sectors in the base year and the 37 input-output sectors in this paper. Through two corresponding matches, the source sectors of the goods can be identified. Additionally, in cases where an HS 8-digit code corresponds to two input-output sectors, the splitting is done accordingly based on the respective proportions. By applying the aforementioned methods to split and match the customs trade import data from 2000 to 2016, the source sectors of the goods can be obtained. The identification rates of the source sector for imported intermediate goods in the customs trade database for each year are shown in Table 5 below:

Table 5 Identification rate of source sector for imported intermediate products in customs trade database

Year	(1) The number of intermediate import records	(2) Import value of intermediate products
2002	99.87%	99.66%
2003	96.93%	85.26%
2004	96.40%	84.22%
2005	95.52%	81.24%
2006	94.14%	77.60%
2007	100.00%	100.00%

2008	97.40%	96.86%
2009	95.31%	94.27%
2010	93.67%	93.89%
2011	92.48%	93.45%
2012	100.00%	100.00%
2013	99.65%	99.79%
2014	96.44%	98.03%
2015	96.27%	98.70%
2016	95.67%	97.79%

5.1.2 Identification of the use sector

As it is assumed that the imported intermediate goods are used in the importing enterprise's own production, the identification of the use sectors of imported products in the customs trade data is also the identification of the sectors to which the importing enterprises belong. In the previous section on initial data processing, information on the input-output sectors of the registered enterprises in the industrial and commercial registration database has been obtained. Since the industrial and commercial registration database contains the enterprises registered in all provinces, cities, autonomous regions, and municipalities directly under the central government from 1949 to 2022, it can be considered to cover the importing enterprises included in the customs trade database. Therefore, by matching the initially processed customs trade database with the initially processed industrial and commercial registration database, the input-output sectors to which importing enterprises belong in the initially processed customs trade data can be obtained.

To improve the matching rate of these two databases, considering the variable information contained in the two databases respectively, the variables used in the matching are the enterprise name, telephone, and email in the customs trade database, and the enterprise name, former name, telephone, and email in the industrial and commercial registration database. Before the matching, the key variables used in the matching of these two initially processed databases are standardized, including removing redundant spaces, unifying bracket formats and letter formats for enterprise names (including former names) and emails, and ensuring consistency in telephone formats. To improve the matching efficiency, the records that can uniquely identify an enterprise with one phone number in the industrial and commercial registration database are saved separately to obtain a corresponding table of phone numbers and input-output sectors. Similarly, the records that can uniquely identify an enterprise with one email in the industrial and commercial registration database are saved separately to obtain a corresponding table of emails and input-output sectors.

The matching strategy for the initially processed customs trade database and industrial and commercial registration database is to sequentially match them based on the enterprise name, telephone, and email. The specific matching steps are: (1) precisely match the two databases based on the enterprise name, and regard enterprises with identical names as the same; (2) for enterprises not matched in the customs database,

match them precisely based on their enterprise names and former names in the industrial and commercial database, regarding enterprises with identical names as the same; (3) for enterprises still not matched in the customs database, precisely match them with the corresponding table of phone numbers and input-output sectors in the industrial and commercial database based on the telephone variable, regarding enterprises with identical phone numbers as the same; (4) for enterprises still not matched in the customs database, precisely match them with the corresponding table of emails and input-output sectors in the industrial and commercial database based on the email variable, regarding enterprises with identical emails as the same; (5) considering that the term "limited company" and "limited liability company" are often used interchangeably in enterprise names, and that "limited company" may be changed to "limited liability company," etc., the enterprise names in the customs database and the enterprise names and former names in the industrial and commercial database containing the following terms are deleted: limited liability, group limited, limited liability, limited, responsibility, equity, company, factory, (group), province, city, county, district, and common punctuation marks. Then the processed enterprise names that were still not matched in the customs database are precisely matched with the processed enterprise names and former names in the industrial and commercial database, and those that matched are regarded as the same enterprise. For the customs data from 2000-2006, the variable information is relatively comprehensive, so matching can be performed according to the steps (1)-(5). However, there are no email and telephone variables in the customs data from 2007-2016, so matching can only be performed according to the steps (1), (2), and (5). The matching results of the initially processed customs trade database and industrial and commercial registration database are shown in columns (1) and (2) of Table 5, and column (3) indicates the proportion of intermediate products import value whose source sector and use sector were both successfully identified. Overall, the success rate of database matching shows an increasing trend, mainly due to the version issues of the databases described in the data source section.

Table 6 Identification rate of source and use sectors for imported intermediate products in customs trade database

Year	(1) The number of intermediate import records	(2) Import value of intermediate	(3) The proportion of import value that both sectors are
2000	45.84%	62.65%	
2001	49.67%	62.47%	
2002	53.75%	64.72%	64.48%
2003	57.81%	65.58%	56.04%
2004	64.52%	73.43%	62.01%
2005	67.77%	76.16%	61.86%
2006	68.82%	73.82%	57.50%
2007	80.49%	88.14%	88.14%
2008	82.82%	89.81%	87.18%
2009	85.24%	90.49%	85.55%
2010	86.69%	90.62%	85.26%

2011	88.05%	91.34%	85.53%
2012	89.21%	91.47%	91.47%
2013	90.62%	91.69%	91.51%
2014	91.91%	91.91%	90.12%
2015	91.90%	91.96%	90.77%
2016	96.89%	96.03%	93.90%

5.1.3 Estimation of intermediate products import shares

When identifying the source sectors and use sectors of products using the above method, the implicit assumption is that the products imported by companies are all used as intermediate inputs for their own production. However, a significant portion of the products imported by the "Wholesale and Retail" and "Transportation, Warehousing, and Postal Services" sectors among the 37 input-output sectors are not used as intermediate inputs for their own production but rather for other sectors. Additionally, due to various reasons, there are significant errors in the import trade data for the "Other Services" sector. Therefore, in estimating the import shares, the import shares for these three use sectors are are determined based on the import proportionality assumption, while the remaining sectors are initially estimated with the original import shares according to the equations $Z_i' = \sum_j Z_{ij}'$ and $\omega_{ij}' = \frac{Z_{ij'}}{Z_{i'}}$. Then, based on the proportions of import shares for each use sector in each source sector, the remaining shares of each source sector after excluding the above three use sectors are allocated. From this, we can obtain the estimation of import shares for the goods sectors.

5.2 Service sectors

Due to the lack of relevant data, it is not possible to identify the use sectors of imported products by identifying the sectors of importing enterprises, as done for imported goods. In this case, the estimation of the IISMs for the service sectors rely on the import proportionality assumption, i.e., assuming that the import of intermediate products is allocated among the use sectors according to the proportion of their overall intermediate use. Therefore, the estimation of the IISMs for the service sectors can be obtained based on the allocation of their products among the use sectors in the competitive input-output tables.

6. Comparison with the IISMs based on import proportionality assumption

This paper mainly focuses on the share of a product's import value assigned to a particular sector, that is, the IISMs Ω, rather than the level of import value. Compared with the methods that rely on the import proportionality assumption to estimate the IISMs, the main improvement of this paper lies in using micro data from China's customs trade database and industrial enterprise registration database to estimate the IISMs of the goods sectors. In order to better demonstrate the differences in estimation results between these two methods, this section compares the IISMs of the goods sectors estimated using the micro data used in this paper with the estimation obtained

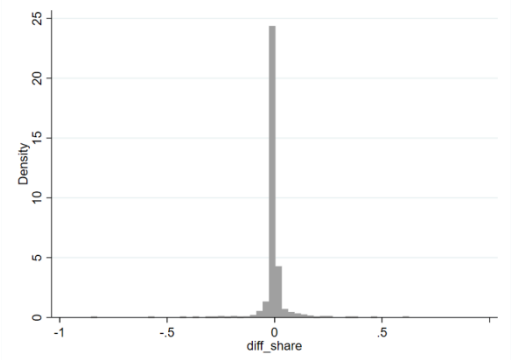
using the import proportionality assumption. Since the National Bureau of Statistics has published input-output tables for benchmark years 2002, 2007, and 2012, the corresponding import matrices can be obtained based on the import proportionality assumption, and the IISMs can be further obtained. Therefore, this section compares the two estimation results for these three years.

First, we calculate the correlation coefficient between the elements in the import coefficient matrices of the goods sectors obtained by two methods for these three years respectively, that is, the correlation coefficient between the import shares of each sector, as shown in Table 6 below. It can be seen that the correlation between the import shares estimated using these two methods shows an increasing trend, with the correlation coefficient reaching 0.796 in 2012, and all correlation coefficients are significant at the 1% level.

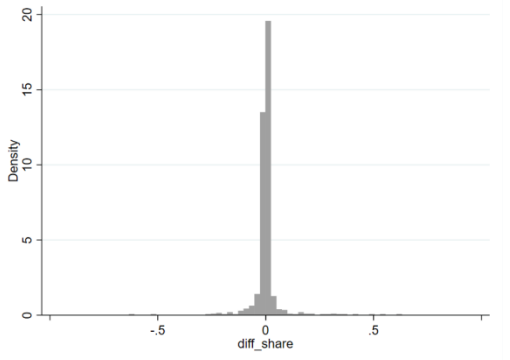
Table 6 The correlation coefficient between the import shares estimated using these two methods.

Year	2002	2007	2012
Correlation coefficient	0.607***	0.699***	0.796***

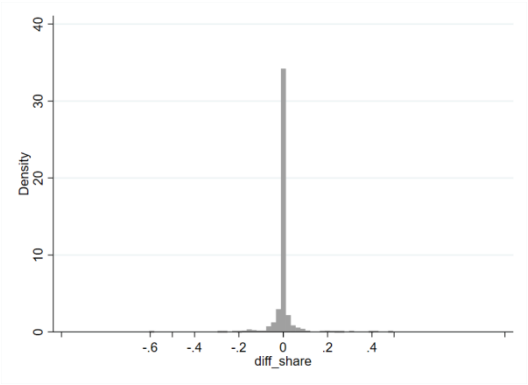
In order to provide a clearer illustration of the differences in estimation results, the histograms of import share differences (estimations based on microdata - estimations based on the proportionality assumption) are plotted as shown in Figure 2 (1)-(3). From the histograms, it can be observed that the differences in import shares of various sectors estimated by these two methods are mostly small. For these three years, the majority of the share differences are within 30 percentage points, and a significant portion of the differences are even smaller, within 10 percentage points. This indicates that a considerable portion of the import allocation to different sectors, as estimated by both methods, are fairly close to each other.



（1）2002



（2）2007



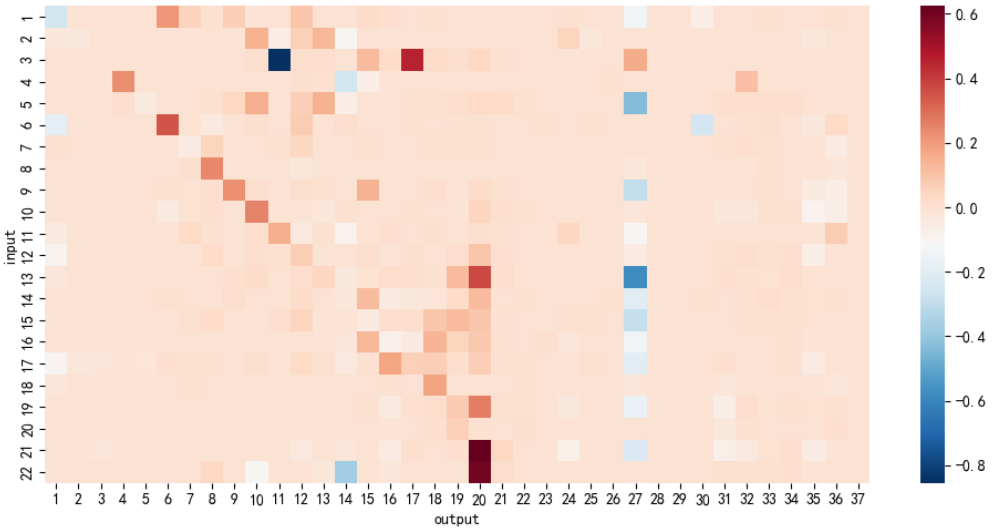
（3）2012

Figure 2 Histogram of differences in estimated import shares in the base year

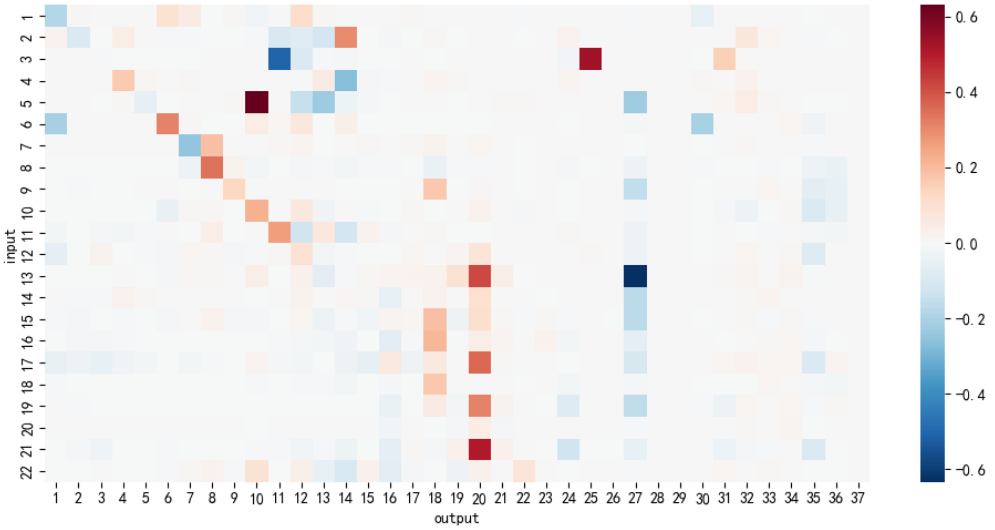
Although the comparison results above indicate that most of the differences in import share estimates are relatively small, there are also some significant disparities within the distribution. In order to demonstrate the differences in results obtained by the two estimation methods more clearly, heatmaps of the share differences for these three years are plotted as shown in Figures 3(1)-(3) below. The calculation method for share differences is based on estimations from microdata minus estimations based on the proportionality assumption. Red squares indicate that the import share estimates obtained from microdata in this paper are higher than those obtained from the proportionality assumption method, while blue squares indicate that the import share estimates obtained from this paper are lower than those obtained from the proportionality assumption method, with darker colors indicating larger differences between the two estimations.

It can be observed that in the heatmaps for 2002 and 2012, most squares are shaded towards red, indicating that a considerable portion of import share estimates for these two years from this study are slightly higher than those obtained from the proportionality assumption method. Conversely, in the heatmap for 2007, most squares are shaded towards blue, suggesting that a considerable portion of import share estimates for this year from this study are lower than those obtained from the proportionality assumption method. Additionally, a notable feature in all three heatmaps is that the squares near the diagonal (representing the first 22 sectors) are mostly shaded with darker colors, indicating significant differences in estimations for these import shares between the two methods.

Looking at the vertical output sectors, particularly the sector of "communication equipment, computers, and other electronic equipment", a considerable portion of squares are shaded towards red, indicating that the proportionality assumption method tends to underestimate the import shares of this sector compared to the microdata-based estimations in this study, suggesting that this sector might rely more heavily on imported goods from other sectors. Conversely, the sector of "construction" exhibits a significant portion of squares shaded towards blue, indicating that the proportionality assumption method tends to overestimate the import shares of this sector compared to the microdata-based estimations in this study, suggesting that this sector might have a weaker dependency on imported goods from other sectors.


input
output
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8

（1）2002


input
output
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

（2）2007

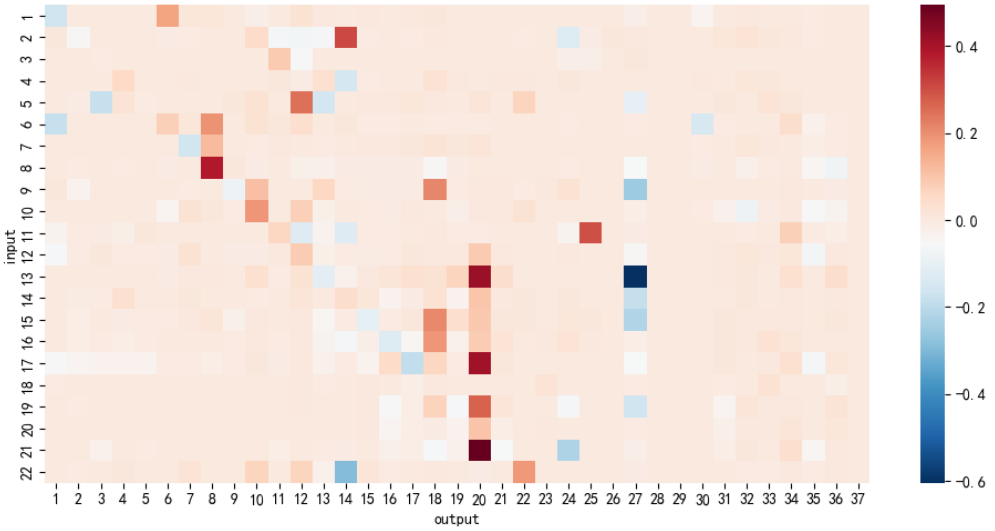

input
output
0.4
0.2
0.0
−0.2
−0.4
−0.6

Figure 2 Heatmap of differences in estimated import shares in the base year

7. Compilation of non-competitive input-output tables

The previous sections have obtained the amounts of imported goods from each sector used for intermediate use, consumption, and capital formation (Z_i', C_i' and K_i'). Based on the amount of imported intermediate products Z_i' and the IISMs estimated in Section 5, the IIFM can be calculated. By embedding it into the competitive input-output tables provided by Zhang et al. (2021) with some necessary adjustments, the corresponding non-competitive input-output tables can be obtained.

Although the IISMs for the goods sectors (22×37) were estimated based on micro data in Section 5, it was found that there was an overestimation in the use of imported intermediate products for some service sectors when embedding them into the competitive input-output table. For example, the estimated usage of imported intermediate products from sector B by sector A was greater than its total usage in the non-competitive table. This overestimation is likely because a large amount of imported intermediate products in the service sectors is not used for their own operational activities. To address this, we adjusted the import shares for the affected service sectors using a proportional assumption. Specifically, we first replaced the original import shares with those obtained using the proportional assumption. Then, based on the sum of the import shares from the same sector equaling one, we adjusted the remaining part of the matrix. By combining the adjusted IISMs for the goods sectors with the IISMs for the service sectors obtained using the proportional assumption, we can get the adjusted IISMs (37×37), denoted as Ω^a.

Using Ω^a and Z_i, we can derive the IIFMs. Subtracting them from the intermediate products flow matrices in the competitive input-output tables yields the domestic intermediate products flow matrices. Similarly, subtracting the estimated values of imported products used for consumption and fixed capital formation from the corresponding values in the competitive input-output tables gives the values of domestic goods used for consumption and fixed capital formation. Based on the competitive input-output table, the initial input part remains unchanged, and changes in inventories and exports also remain unchanged (excluding imports in inventory changes and re-exports). By making the above adjustments, decompositions, and integrations for the remaining parts, we obtain the non-competitive input-output table.

8. Conclusion

Currently, GVCs, which are mainly characterized by production fragmentation and trade integration, have become the primary form of global production, with intermediate goods trade playing an increasingly important role. A large amount of literature has studied the new type of international trade under GVCs and its impact, but all these studies are limited by data quality to some extent. Existing estimates of import matrices rely on the import proportionality assumption, which is an overly simplified assumption that hardly reflects the reality accurately, greatly affecting the reliability of relevant research. To address this issue, researchers have started using

more detailed micro data to directly estimate sectoral import matrices. However, existing methods are based on special data and are practically infeasible to apply and generalize. Moreover, the data used are often sampled, resulting in low estimation accuracy.

This paper proposes a new method for estimating import matrices that does not rely on the import proportionality assumption. Based on this method, this paper also compiles China's time series non-competitive input-output tables for 37 sectors during the period of 2000-2016. The main feature of this method is that it is based on real micro data and combines machine learning algorithms to improve estimation accuracy. Without depending on the proportionality assumption, it estimates the proportion of imported goods used for intermediate use, consumption, and fixed capital formation by using BEC classification. Then based on the total imports of various products in the competitive input-output tables, we can obtain the estimation of the values of imported goods used for these three purposes. For the estimation of import matrices, the sectors are divided into goods sectors and service sectors, each using different estimation methods. For goods sectors, customs trade data are used as the basis. First, they are matched with the concordance table between HS 8-digit code and the input-output sector to identify the source sector of the product. Then, they are matched with the enterprise registration database to identify the sector to which the enterprise belongs, which represents the use sector of the product. By identifying the source and use sectors of imported products, the allocation of imported goods among sectors can be determined, and the IISMs can be estimated through simple calculations. For service sectors, the estimation of import coefficient matrices can be directly obtained based on the import proportionality assumption. After obtaining the IISMs, based on the estimated total amount of each imported product used fro intermediate use, the import matrix can be obtained. By embedding the import matrix into the competitive input-output tables and making necessary adjustments, we can obtain the non-competitive input-output tables. By comparing the IISMs estimated using the micro data used in this paper with the estimation obtained using the import proportionality assumption method, we find that, overall, the differences in most import share estimates between the two methods are not significant. However, there are also some sectors with large differences in share estimates, exceeding 50 percentage points.

The enterprise registration data and customs trade data used in this paper are statistically collected by most countries worldwide. Therefore, this estimation method not only improves the accuracy and timeliness of import matrix estimation but can also be popularized and applied to other economies. The use of machine learning methods to identify the sectors to which enterprises belong based on their business scope can also be extended to other economies. In addition, further improvements can be made to this method. According to the mentioned issue of database versions, if the version of the enterprise registration databases that is closer to the matching year is used when matching the customs trade database and enterprise registration database in each year, the matching rate will be further improved, thus improving the accuracy of the estimation .

Reference

[1] Ahn J B, Khandelwal A K, Wei S J. The role of intermediaries in facilitating trade[J]. Journal of International Economics, 2011, 84(1): 73-85.

[2] Alfaro L, Chor D, Antras P, et al. Internalizing global value chains: A firm-level analysis[J]. Journal of Political Economy, 2019, 127(2): 508-559.

[3] Antràs P, Chor D, Fally T, et al. Measuring the upstreamness of production and trade flows[J]. American Economic Review, 2012, 102(3): 412-416.

[4] Antràs P, De Gortari A. On the geography of global value chains[J]. Econometrica, 2020, 88(4): 1553-1598.

[5] Caliendo L, Parro F. Estimates of the Trade and Welfare Effects of NAFTA[J]. The Review of Economic Studies, 2015, 82(1): 1-44.

[6] Carrico C, Corong E, van der Mensbrugghe D. The GTAP version 10A multi-region input output (MRIO) data base[R]. Center for Global Trade Analysis, Department of Agricultural Economics, Purdue University, 2020.

[7] Chen Q, Chen X, Pei J, et al. Estimating domestic content in China’s exports: Accounting for a dual-trade regime[J]. Economic Modelling, 2020, 89: 43-54.

[8] Dean J M, Fung K C, Wang Z. Measuring vertical specialization: The case of China[J]. Review of International Economics, 2011, 19(4): 609-625.

[9] Dietzenbacher E, Los B, Stehrer R, et al. The construction of world input–output tables in the WIOD project[J]. Economic systems research, 2013, 25(1): 71-98.

[10] Ertur C, Koch W. Growth, technological interdependence and spatial externalities: theory and evidence[J]. Journal of applied econometrics, 2007, 22(6): 1033-1062.

[11] Feenstra R C, Hanson G H. Globalization, outsourcing, and wage inequality[J]. 1996.

[12] Feenstra R C, Hanson G H. The impact of outsourcing and high-technology capital on wages: estimates for the United States, 1979–1990[J]. The quarterly journal of economics, 1999, 114(3): 907-940.

[13] Feenstra R C, Jensen J B. Evaluating estimates of materials offshoring from US manufacturing[J]. Economics Letters, 2012, 117(1): 170-173.

[14] Houseman S, Kurz C, Lengermann P, et al. Offshoring bias in US manufacturing[J]. Journal of Economic Perspectives, 2011, 25(2): 111-132.

[15] Hummels D, Ishii J, Yi K M. The nature and growth of vertical specialization in world trade[J]. Journal of international Economics, 2001, 54(1): 75-96.

[16] Johnson R C, Noguera G. Accounting for intermediates: Production sharing and trade in value added[J]. Journal of international Economics, 2012, 86(2): 224-236.

[17] Kee H. L. and H. Tang. Domestic Value Added in Exports: Theory and Firm Evidence from China[J]. American Economic Review, 2016, 106(6): 1402-1436.

[18] Koopman R, Wang Z, Wei S J. Tracing value-added and double counting in gross exports[J]. American economic review, 2014, 104(2): 459-494.

[19] National Research Council (committee report), 2006. Analyzing the US content of imports and the foreign content of exports, Washington, DC, National Academies Press.

[20] Oosterhaven J, Stelder D, Inomata S. Estimating international interindustry

linkages: Non-survey simulations of the Asian-Pacific economy[J]. Economic Systems Research, 2008, 20(4): 395-414.
[21] Puzzello L. A proportionality assumption and measurement biases in the factor content of trade[J]. Journal of International Economics, 2012, 87(1): 105-111.
[22] Timmer M P, Los B, Stehrer R, et al. Fragmentation, incomes and jobs: an analysis of European competitiveness[J]. Economic policy, 2013, 28(76): 613-661.
[23] Timmer M P, Dietzenbacher E, Los B, et al. An illustrated user guide to the world input – output database: the case of global automotive production[J]. Review of International Economics, 2015, 23(3): 575-605.
[24] Winkler D, Milberg W. Bias in the'Proportionality Assumption'Used in the Measurement of Offshoring[J]. World Economics-Abingdon, 2012, 13(4): 39.
[25] Yu M. Processing trade, tariff reductions and firm productivity: Evidence from Chinese firms[J]. The Economic Journal, 2015, 125(585): 943-988.
[26] Zhang H, Xia M, Su R, et al. The Compilation of the Time Series Input-output Tables in China: 1981-2018[J].Statistical Research,2021,38(11):3-23.