Chapter 2
Conversion of Trade Commodity Classification SITC to IO24 Sector Classification: Processing Conversion Error Data

KUROKO Masato

Before creating a world trade matrix based on the IO20 industrial category codes (hereafter IO20), we need to convert the SITC product category codes (hereafter the SITC) used in the original AID-XT data to the IO20. Here I will provide an outline of conversion processing and specific processing procedures, with an emphasis on the processing of conversion error data. The first conversion method I will discuss is the average distribution formula, followed by the value-weighted distribution formula.

1. Conversion According to the Average Distribution Formula

In order to convert the SITC into the IO20, we first search the SITC/IO20 conversion tables (hereafter, conversion tables), using the SITC as the key, and when we find a relevant SITC in the table, we assign its IO20 to the corresponding IO20. The conversion tables referred to here are table files, which give correspondences between one four- or five-digit SITC and one IO20. If there is a relevant SITC, it is written as an IO20 of output data. The data resulting from this conversion process are the pass data.

Next, a problem arises when there are errors resulting from searching the SITC/IO20 conversion tables using the SITC as the key, so our conversions according to the average distribution method use the conversion table to generate an extended conversion table, which we then use as the basis for conversion. What we mean by an extended conversion table is a table that notes the frequency ratios of the correlations between the SITC and the IO20 based on the original conversion tables.

The average distribution formula uses these tables to distribute the correlations equitably according to their frequency ratios. The specific procedures are described below.

[1] Generating the Extended Conversion Tables: We first create a table that registers all the patterns, dropping the last digits of the SITC's on the conversion table one at a time. For example, if the SITC in the original table is 05461, as in Fig. 6 in this document, it is written out separately as 0, 05, 054, 0546, and 05461. Since this will be used later as extended version of the original conversion table, we will refer to it as the extended conversion table. Next, we take the IO20 from the extended conversion table and repeatedly write them out in a horizontal array of 21. Then we add 1 to the corresponding column of IO20 and the 21st column (see Fig. 8 in this document). After processing all the conversion data this way, we tally up columns 01 to 20 and the total column for each SITC. Fig. 9 contains an example of the results of processing an error data (4) (Fig. 5) by this method.

[2] Conversion According to the Original Conversion Tables: Now we come to the actual conversion. The first step is to take the original, not the extended, conversion table and the AID-XT data and conduct a table lookup. The output, as shown in Fig. 1, is pass data and unconvertible error data.

[3] The First Conversion According to the Extended Conversion Tables: Next, we search the extended conversion tables created in [1] for the error data generated in [2]. At this point the SITC from AID-XT data are still in their original form, without having undergone the processing that removes their final digit. On the other hand, the extended conversion tables make it possible to process the SITC on the original tables and search them at all levels of digits. In cases such as those shown in Example (1) (Fig. 2) and Example (3) (Fig. 4), this process allows the AID-XT data to be converted and output as pass data.

On the other hand, in cases such as those shown in Example (2) (Fig. 3) and Example (4) (Fig. 5), they will not match unless the SITC in the AID-XT data are changed, and in this case, they are output as error data.

[4] The Second and Subsequent Processing Sequences According to the Extended Conversion Tables: In the previous steps, we have done the matching without processing the SITC from the AID-XT data, but in this step, we process these SITC and search it with the extended conversion tables. We repeat a "do while" loop while removing one digit at a time from the end of the SITC that is read in from the AID-XT data until we achieve a match. Since even in the worst case, we can expect that there will definitely be an SITC at the 1-digit level that is appropriate for the extended conversion table and that we can achieve a match, this type of processing outputs all the AID-XT data as pass data converted to IO20.

[5] Sorting and Summarizing the Pass data: The pass data output to this point are sorted and summarized.

A weak point of the average distribution formula that we might mention is that sometimes the results cannot be made to conform to the actual records of amounts and volumes of the commodities.

An advantage is that in the end, that no completely non-convertible error data is generated.

2. Conversion According to the Value-Weighted Distribution Formula

The value-weighted distribution formula is a conversion method proposed as an attempt to correct the weak point of the average distribution formula, namely the possibility of a distribution that does not match the actual records. The value-weighted distribution formula and the average distribution formula differ in the way they process the AID-XT data that could not be converted in the original conversion table.

In the average distribution conversion method, we basically expand the conversion tables and eliminate the error. On the other hand, in the value-weighted distribution conversion formula, we first use the original conversion tables and convert the AID-XT data, as in the other formula, but after that, we generate conversion tables independently from the pass data, and then we weight and distribute them, not by the frequency of the IO20 corresponding to each SITC, but by the aggregate of the relevant values. These independently generated conversion tables differ from the average distribution in that the reporting country, the direction of trade, the year reported, and the SITC at every level of digits serve

as keys.

The specific procedure is as follows:

[1] Conversion According to the Original Conversion Tables: We first convert the data according to the original conversion tables. This is exactly the same as the first conversion in [2] for the conversion method based on the average distribution.

[2] Generation of the Extended Conversion Tables from the Pass data: Next, we generate conversion tables independently from data that has been output as pass data. (Hereafter, this will be referred to as the extended conversion tables). The extended conversion tables are in the following format: Reporting country, Direction of trade, Year reported, SITC at every level of digits, IO20 (20 arrays).

First, we select only the items corresponding to the format above from among the items in the pass data (Fig. 10 in this document). Here RC stands for "reporting country," DT stands for "direction of trade," Y for "year reported" (2 digits based on the Western calendar), and V for "value." Note that the value differs from the amounts in the actual data.

Next, we output the data for each SITC for every level of digits and sort them in order of the keys (Fig. 11 in this document).

The next step is to summarize the data that have the identical keys on one line. Placing the IO20 into an array, the proportion of the value of each IO20 is stored as a percentage of the total value that it accounts for within each key. This completes the generation of the extended conversion tables (Fig. 12 in this document).

[3] The First Conversion with the Extended Conversion Tables: Now we come to the actual conversion. The AID-XT data that was identified as error data in [1] and the extended conversion tables generated in [2] are read in and matched. At this point, the SITC in the AID-XT data have not yet been processed and are still in their original state.

[4] The Second and Subsequent Conversions with the Extended Conversion Tables: From here on, we match the SITC from the AID-XT data with the extended conversion tables, dropping final digit at a time. This is functionally equivalent to the second and subsequent conversions according to the extended conversion tables in the average distribution formula in [4]. But since this is file matching, as similar to that in [3], the program that drops the final digits of the SITC in the AID-XT data, the sorting with summarizing, and the matching program are all consolidated in one process. This process is repeated until no more pass data are output. If we assume the case of the SITC, which have a maximum of 5 digits, the worst-case scenario will require 4 repetitions of [4] to reach the 1-digit level.

[5] Final Conversion of the Error Data and the Sorting and Compilation of the Pass data: Unlike the average distribution formula, the value-weighted distribution can leave AID-XT data that cannot be converted, even if the SITC is at the 1-digit level. This occurs when both the following conditions are present:

The first condition is when an error occurs in Step [1]. The other condition is when the data in which the reporting country, the direction of trade, the year reported, and the SITC at the 1-digit level are equal to the final error data does not pass the Step [1]. When these conditions occur together, we end up with non-convertible error data. Being non-convertible, these data are assigned to the IO20 code 99 "Miscellaneous."

At the end, the final error data and the pass data are sorted and summarized together.

The advantage of the value-weighted distribution formula is that since the distribution is conducted on the basis of the actual records of the amounts, the appropriateness of the distribution may be better than that for the average distribution formula. Note, however, that we are referring to cases in which there is a great deal of pass data in Procedure [1] and valid extended conversion tables are generated in [2].

Conversely, the disadvantage of this formula is that it may not be possible to get rid of all the error data in the end. Since a valid extended conversion table cannot be generated to include most types of keys in [2] if there is little pass data in Procedure [1], we end up with a lot of error data. In such cases, we cannot make an absolute statement that this method is generally more appropriate than the average distribution formula. It may be said that the appropriateness of the value-weighted distribution formula depends in some sense on the completeness of the original conversion tables.