Chapter 1

Evaluation and Correction for Consistency of UN Comtrade Data

NODA Yosuke

Introduction

The commodity classifications used in the UN Comtrade data formulated by the International Trade Statistics Section of the UN Statistics Division contain commodity classification codes of all digit levels organized hierarchically. Because the totals of transaction values for the lower digit levels do not necessarily match totals for the corresponding upper digit levels however, it is known that there is cause to examine the consistency of this trade data.

One purpose of this chapter is to provide an explanation of the method enabling evaluation and correction to the fullest extent possible of trade data drawn from the UN Comtrade data, in particular in relation to commodity classifications, as revised in 2005 by Noda and Fukao. Another purpose is to formularize the concepts behind digit-level classification codes, mdcc classification codes, and the evaluation and correction of consistency, and to employ these formulas to enable organization and expression of these procedures.

1. Commodity Classification of Trade Data

UN Comtrade data contains transaction value data for each of the digit-level classification codes which are hierarchically structured to form the commodity classifications. The set of commodity classifications formed from k digit level classification codes will be termed $C(k) = \{ c_1(k), \cdots, c_{m(k)}(k) \}$, and partner countries will be termed $P = \{World, p_1, \cdots, p_n\}$. World for partner country will be expressed as World, the number of partner countries as n, and the number of n, k digit level commodity classification codes as $m(k)$, where $m(0) = 0$. For SITC commodity classifications, $k = 1, \cdots, 5$, and for HS commodity classifications, $k = 2,4,6$. Transaction value data obtained as trade data can be expressed as

(1-1) $v_{rc,d,y}(c_i(k), p_j)$

for the commodity classification codes expressed as k digit level commodity codes, $c_i(k)$, and partner countries p_j , for each reporting country rc, year y, and direction of trade d, where $c_i(k) \in C(k)$ and $p_j \in P$. When there is no confusion, equation (1-1) can be expressed simply as $v(c_i(k), j)$ or $v(c, j)$. If the total value of commodities in the commodity classifications is expressed as T and world for partner country is expressed as W, $v(T, W)$ is simultaneously total value of commodities and transaction value for partner country world.

The set of all commodity classification codes will be termed Ω. For each of the SITC revisions, the elements making up Ω are *Total*, incorporating all commoditys, and the set of hierarchically-structured digit-level commodity codes, from the 1 digit level

commodity codes to the 5 digit level commodity codes, and

$$\Omega = \{Total, C(k)\ k = 1 \cdots 5\}$$

where $Total = c_1(k) \cup \cdots \cup c_{m(k)}(k)$ for $k = 1 \cdots 4$. The organization of the 5 digit level classification codes differs from that of the other classification codes, and is not hierarchical. If the discriminant function expressing the fact that commodity classification code c is composed of Ω elements is termed ξ, for $c \in \Omega$, $\xi(c) = 1$, and for $c \notin \Omega$, $\xi(c) = 0$. The set of 1 digit numbers from 0 to 9 is termed $A_S = \{0,1,\cdots,9\}$. The set of k digit level commodity codes for each SITC revision can be expressed as

$$\begin{aligned} & C(k) = \{\omega \mid \omega \in (C(k-1) \times A_S), \\ & \xi(\omega) = 1\} \end{aligned} \tag{1-2}$$

for $k = 1, \cdots, 5$, where $C(0) = \phi$ and $C(k-1) \times A_S$ is the Cartesian product of $C(k-1)$ and A_S. The elements of $C(k-1) \times A_S$ taken as candidates for lower digit level commodity codes for $C(k-1)$ are not necessarily confined to the set encompassed by Ω. Because of this, ξ is used to limit elements to those encompassed by Ω in equation (1-2).

2. *mdcc* code as the most detailed code

When the SITC revisions are employed as the commodity codes, the set of $k+1$ digit level classification codes $D_S(c)$ for which the k digit level is $c \in C(k)$ is defined as

$$\begin{aligned} & D_S(c) = \{\omega \mid c \in C(k), \omega \in C(k+1), \\ & \eta_k(\omega) = c\} \end{aligned} \tag{2-1}$$

for $k = 1, \cdots, 5$,where $D_S(c) \subset C(k+1)$ and naturally, $\eta_k(D_S(c)) = c$ for $c \in C(k)$. Using equation (2-1), the entire set of $k+1$ digit level classification codes can be expressed. This becomes

$$C(k+1) = \{\omega \mid \forall c \in C(k), \omega \in D_S(c)\} \tag{2-2}$$

for $k = 0, \cdots, 4$, whetre c in $\forall c \in C(k)$ means all elements included in $C(k)$. Because the set of $D_S(c)$ for $c \in C(k)$ is Ω for the entire set of commodity classification codes,

$$\begin{aligned} \Omega \setminus \{Total\} &= C(1) \cup \cdots \cup C(5) \\ &= \{\omega \mid \omega \in D_S(c), \forall c \in C(k), \\ & k = 0 \cdots 4\}. \end{aligned} \tag{2-3}$$

Separately to the hierarchically-structured commodity classification codes, detailed codes for each reporting country, direction of trade, and year with a transaction value greater than zero in trade data but which do not have classification codes in the lower-digit strata are termed the most detailed classification codes, or *mdcc*. With regard to partner countries, because $v(c,W) \geq v(c,j)$ for $j \in P \setminus \{World\}$, general rule is preserved even when the transaction value is taken as $v(c,W)$ for partner country world. Because the set of *mdcc* classification codes M_S are the set of most detailed classification codes when transaction value is considered,

$$\begin{aligned} & M_S = \{c \mid c \in C(k), k = 1 \cdots 5, \\ & v(c,W) \neq 0, \\ & \forall \omega \in D_S(c), v(\omega,W) = 0\}. \end{aligned} \tag{2-4}$$

where $v(\phi,W) = 0$.

For each SITC revision, the *mdcc* can be expressed by equation (2-4), and for HS Original and the HS revisions, they can be expressed by equation (2-6). When an SITC revision is employed, the *mdcc* commodity codes included in the k digit level classification codes can be expressed as M_k in equation (2-8), and this becomes

$$\eta_k(N_{k+1} \cap M_k) = \phi$$

for $k = 1 \cdots 4$, based on equation (2-12). Establishing the condition $c \in M_k$ for k digit level classification codes ensures that no elements for which the k digit level is c exist in the $k+1$ digit level classification codes. The same holds true for HS Original and revisions. Equation,

$$\sum_{k=1}^{5} \sum_{c \in M_k} v(c(k),W) = v(T,W) \tag{2-17}$$

expresses the fact that the total transaction value for

mdcc matches the total transaction value for commodities in trade data for which consistency between partner countries and commodity classifications has been ensured. It is necessary to be aware here that the *mdcc* are not necessarily hierarchically structured.

3. Evaluation for Consistency of Trade Data

In cases in which the trade data is grouped into *mdcc* but consistency has not been ensured, equation (3-11) states that at least one classification code for which $\alpha(c_i) \neq 0$ must exist among the k digit level classification codes for which $c_i \in C(k)$. When consistency has been ensured, $\alpha(c_i)$ becomes 0, and the general equation for comparison of transaction values between digit-level classification codes can be expressed as equation (3-11).

If c in equation (3-11) is the Total for all commodities, when the transaction value $v(T,W)$ and the difference between $C(1)$, the set of 1 digit level classification codes, and the sum of transaction values do not match, because $N_1 = C(1)$, equation (3-17) can be generated using equation (3-14). When equation (3-11) is substituted for the second term on the right-hand side of equation (2-11), equation (2-15) generates equation,

$$\begin{aligned} \sum_{c\in N_k} v(c(k),W) &= \sum_{c\in M_k} v(c(k),W) \\ &+ \sum_{c\in N_{k+1}} v(c(k+1),W) \\ &+ \sum_{c\in N_k \setminus M_k} \alpha(c(k)) \end{aligned} \tag{3-18}$$

for $k = 1 \cdots 4$. Using $k = 1$ in equation (3-18) and substituting this in equation (3-17), we obtain

$$\begin{aligned} v(T,W) &= \sum_{c\in M_1} v(c(1),W) + \sum_{c\in N_2} v(c(2),W) \\ &+ \sum_{c\in N_1 \setminus M_1} \alpha(c(1)) + \alpha(c_\bullet(0)) \end{aligned}$$

Repeating this operation for values of *k* to be from 2 to 4 generates equation,

$$\begin{aligned} v(T,W) &= \sum_{k=1}^{5} \sum_{c\in M_k} v(c(k),W) + \alpha(c_\bullet(0)) \\ &+ \sum_{k=1}^{4} \sum_{c\in N_k \setminus M_k} \alpha(c(k)) \end{aligned} \tag{3-19}$$

For trade data for which consistency has not been ensured, equation (3-19) can be used to express the commodity total as the *mdcc* for each digit-level classification code and the sum of error. The total of the absolute error of the digit-level classification codes in the *mdcc* can be expressed as equation (3-20). The error for 1 digit level classification codes is

$$\alpha(c_\bullet(1) = v(T,W) - v(c_\bullet(1),W).$$

The error for *k* digit level classification codes for $k = 1 \cdots 4$ is expressed as the third term on the right-hand side of equation (3-18). The error for trade data for which consistency has been ensured, as shown by equation (3-20), is 0, and this matches equation (2-17). Equation (3-19) is therefore the generic formula for evaluation of error in *mdcc*.

For a trade matrix based on hierarchical structured digit-level commodity classifications and separate partner countries, when the reporting countries, years, and direction of trade are fixed, the error originating with commodity classifications is expressed by equation ,

$$e_c(k) + e_{c,p}(k) = v(T,W) - v(c_\bullet(k),W) \tag{3-9}$$

the error originating with partner countries is expressed by equation ,

$$e_p(k) + e_{c,p}(k) = v(T,W) - v(T,\bullet) \tag{3-10}$$

and the total error including commodity classifications and partner countries is expressed by equation,

$$e(k) = e_c(k) + e_p(k) + e_{c,p}(k)$$

and this equation induces

$$e(k) = v(T,W) - v(c_\bullet(k),\bullet). \tag{3-8}$$

The consistency evaluation tables compile all these forms of error. Tables 7 to 12 are the consistency evaluation tables for the *k* digit level classification codes.

4. Correlation of Trade Data

In the case of trade data for which consistency has been ensured for partner countries, when units of quantity are identical, two correction criteria, using absolute error and relative error, are available to enable the consistency of commodity classifications to be maintained. Signed absolute error is expressed by equation (3-11), and relative error is expressed by equation (3-12). If the correction criteria for absolute error is termed α^*, when the correction of absolute error in commodity classification code c satisfies equation (4-2) and the correction of relative error satisfies equation (4-6), simultaneous addition of $(c \times m) \in C(k+1)$ and its transaction value $\alpha(c)$ to the items to be corrected for the $k+1$ digit level classification code gives equation

(4-5) $$D_S(c)^* = D_S(c) \cup (c \times m) .$$

UN Comtrade data cannot be corrected when the correction criteria are unsatisfied. Because consistency is generally not guaranteed for partner countries, when partner countries and commodity classifications have been corrected, the two categories are finally corrected together, because they contain error that has been rounded off when absolute error, α^*, did not equal 0.