Chapter 3

Evaluation to Consistency of Trade Statistics with Reversion to Commodity Classification

NODA Yosuke

This chapter will introduce a method for evaluating the consistency that occur in time-series connections as a result of revisions to the commodity classification systems when using global trade statistics data. Its additional purpose is to study the results obtained by applying this method to the data for Japan in the OECD trade statistics. The trade statistics are set as time-series data for the period 1962 to 1999 in the form of (1) the series that takes only the SITC numbers as the classification codes, without considering differences due to SITC revisions, (2) the series grouped together in Chapter 1, and (3) the series created in Chapter 1 on the basis of distributed weight. For each of these series, we examine whether or not structural changes occurred around the years when commodity classifications were revised.

In general, the point in time when structural change occurs in a time-series is called a changing point. When a classification code or a commodity group has a changing point that coincides in time with a revision of commodity classifications, this does not necessarily signify the existence of an inconsistency, but if there is an inconsistency, then it is possible that a changing point will coincide with the time of commodity classification revision. The coinciding of the time of revision of commodity classifications with a changing point can be considered a necessary condition for evaluation of the inconsistency that accompanies the conversion of commodity classifications. The changes in transaction values that arise when revision takes place in this way are treated in this chapter as structural changes in the time series. The chapter applies two methods to make a determination with respect to these structural changes. One is the regression model, which divides the data close to the time when the commodity classifications are revised. The other is the posterior distribution method applied when the changing pointe is taken as a parameter.

The increase in mean value of time-series data is often accompanied by increased scatter or conspicuously uneven distribution. This kind of data cannot be processed very well by statistical analysis that presupposes a distribution occurring in accordance with normal distribution. This kind of data can be handled with Box-Cox transformation, which is an ordinary data conversion used to obtain data that maintain a more or less fixed fluctuation.

Take time-series $\{y_1 \cdots y_n\}$ to represent the value of trade transactions. We will not use the entire range of time-series data that

exists with respect to the revisions from SITC-R1 to SITC-R2, but instead will take only the annual data from 1962 to 1987 that affects those revisions, and will divide this segment of data into two parts. This division consists of the assumption that the period of applicability for SITC-R1 is [62,m] and that for SITC-R2 is [m+1,87]. A linear regression equation of the following kind is then applied to each of the period divisions. Taking the parameters to be $\{\alpha_i, \beta_i\}$ $i=1,2$ the linear regression equation will be expressed in divided form as follows:

$$(6) \quad y_i = \begin{cases} \alpha_1 + \beta_1 x_i + e_i & i = 1,\cdots,m \\ \alpha_2 + \beta_2 x_i + e_i & i = m+1,\cdots,n \end{cases}$$

where assuming that the disturbance term e_i occurs in accordance with normal distribution, then $e_i \sim N(0,\sigma^2)$ $i=1\cdots n$. As the time of the division, m is defined to move through the range of years from 1963 to 1985. With respect to the revisions from SITC-R2 to SITC-R3, we will use only the data from 1978, which affects the revision, to 1999, for which the data exist, and will divide this segment of data into two parts. This division consists of the assumption that the period of applicability for SITC-R2 is [78,m] and that for SITC-R3 is [m+1,93]. Let us apply a linear regression equation, as with equation (6), to each of the period divisions. The range of movement for the point of division m is defined as 1979 to 1997. The linear regression equation (6), which expressed the segment division, is expressed as a matrix as follows:

$$(7) \qquad y = X\theta + e$$

Here, y is the n-dimensional vector of the observed value, and e is the n-dimensional vector of the disturbance term, so that $e \sim N(0,\sigma^2 I)$. X is the matrix $n \times 4$ while θ is 4 dimensional vector. X changes with division at time m.

$$X = \begin{bmatrix} 1 & x_1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_m & 0 & 0 \\ 0 & 0 & 1 & x_{m+1} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & x_n \end{bmatrix} \quad \theta = \begin{bmatrix} \alpha_1 \\ \beta_1 \\ \alpha_2 \\ \beta_2 \end{bmatrix}$$

The maximum likelihood estimator for θ and σ^2 can be obtained by a regular simultaneous equation in which partial differentiation of the log-likelihood function by θ and σ^2 yields 0. X changes with division at time m, so that the maximum log-likelihood is a function of m. Consequently, the AIC in the equation (7) model is expressed as a function of m as follows:

$$(13) \qquad \mathrm{AIC(m)} = n\{\log(2\pi\hat{\sigma}^2) + 1\} + 10$$

where

$$\hat{\sigma}^2 = (y-\hat{y})'(y-\hat{y})/n.$$

The m with the minimum AIC(m) will be the estimated value of the changing pointe.

The parameters used in a linear regression equation are θ, m, and σ^2, but in this equation they are defined as unknown fixed values. Instead, we will define these parameters as random variables. The respective density functions of prior distributions for the random variables θ, m, and σ^2 will be:

$$f(m) = \begin{cases} \dfrac{1}{n-3} & m = 2,\cdots,n-2 \\ 0 & otherwise \end{cases}$$

$$f(\theta) \propto c$$

$$f(\sigma^2) \propto \frac{1}{\sigma^2}$$

where $\propto$ represents a proportional connection, and c is a constant. Accordingly, the marginal density function of the posterior distribu-

tion for m will be:

$$f(m \mid y) = c_3 \mid X'X \mid^{-\frac{1}{2}} \{(y-\hat{y})'(y-y')\}^{-\frac{n-4}{2}} \quad (15)$$

where $m = 2, \cdots, n-2$ and c_3 is a constant.

The SITC has undergone sequential revisions from SITC-R1 to SITC-R3, but it will happen sometimes that when there is no need for examination of the detailed commodity classifications, we will use the 1 digit level or 2 digits level of the commodity classification SITC without considering the differences that arise from the commodity classification systems. This is founded on the hypothesis that, although trade transaction values directly reflect differences in classifications at the level of the basic items (the commodity classification codes that are expressed by the most detailed individual classifications), the differences from basic items will cancel each other out at the higher levels, such as the 1 digit level or 2 digits level, so that they do not appear on the surface. This is frequently used as a handy, simple method.

Table 1 in Part 2 presents the commodity classification codes and their designations at the 1,2, and 3 digits levels of the SITC for each of its revisions. At the 1 digit level in this table, most of the classifications show more or less similar commodities lined up side by side. The differences are evident, however, when we go down to the 2 digits and 3 digits levels. As an extreme example of the relationship between SITC-R1 and SITC-R2, there are classification codes in SITC-R2 that do not exist in SITC-R1, and the converse is also the case.

Table 4 in Part 2 shows the results obtained from the series when only the identical 3 digit numbers are used as classification codes, without considering the differences arising due to SITC revisions. The correspondences in "SITC-R1 : SITC-R2" and "Import" are between SITC-R1 and SITC-R2, and "2 digits level of SITC," the import series, shows the 2 digits level. The X-symbols in this table signify places where the two models provide the same results, while the asterisks (*) show where only one of the results matches. The crosshatch symbols (#) mark where there are no data for the entire period from 1962 to 1987, due to defective figures in the data caused either by absence of classification codes in the commodity classification system or by low transaction volumes. Where D is a period (.) and also at the same time an X or an asterisk, it is a 2 digits classification code for which the revision year and the changing point coincide. Where D is a crosshatch symbol and also at the same time an X or an asterisk, it indicates that the classification code does not exist in one or other.

Table 2 summarizes Table 4 in Part 2. The correspondences between SITC-R1 and SITC-R2 show that approximately 60% of the structural change in both exports and imports is to be found in the revision year. The correspondences between SITC-R1 and SITC-R2 show that imports account for approximately 50% of this, while exports account for approximately 35%.

Table 3 presents the results from commodity group series based on groupings of correspondences at the 3 digits level. Of SITC The commodity groups were created from UN commodity tables, so they actually should not

show any structural changes in the revision year. It is necessary to note, however, that there are commodity groups that undergo structural change despite having type 1 correspondences.

Section 3 in Chapter 1 estimates each SITC series as time-series data for the period from 1962 to 1999 based on the distributed weight with respect to commodity classification codes at the 3 digits level of SITC in Japan's trade statistics. In order to evaluate the estimated weight, a comparison series was created by uniforml distribution of the SITC 3 digits level.

Table 5 in Part 2 presents the results from evaluation of each SITC series by means of the two models explained earlier. In this table, *NN* signifies series created by distributed weight estimated by neural networking, and *ED* signifies series created from weights that are equally distributed. The table first shows the import series for SITC-R1. The "1 digit level of SITC" presents results from the study of imports at the 1-digit level of SITC, the "2 digits level of SITC" those from 2-digit level imports, and the "3 digits level of SITC" those from 3-digit level imports. Next, each series of SITC-R1 exports is shown in the same way, and each series of imports and exports for SITC-R2 and SITC-R3 is then also shown in the same way.

To summarize the results from Table 4, the series for which *NN* was judged to be the better distribution method at the 3 digits level, according to the correspondence between SITC-R1 and SITC-R2, are both the import and export series of SITC-R1 and both the import and export series of SITC-R2. However, *ED* is judged to be the better distribution method for both the import and export series of SITC-R3.

To reiterate, the existence of classification codes and commodity groups for which the points of change coincide with the revision years does not necessarily signify that there is a correspondence inconsistency that occurred with the commodity classification revisions. However, if there is an inconsistency, then it is possible that a changing point will coincide with the year of commodity classification revision. From such a perspective, the methods introduced in this chapter do no more than identify candidate classification codes for which structural changes can be seen as occurring in the revision year according to certain criteria for correspondence among the large number of classification codes that exist. Consequently, study of the matter ultimately requires us to compare examples of commodity items for the corresponding individual classification codes.