SUMMARY

Preface

Rapid globalization of the world economy has resulted in significant changes in the circumstances surrounding world trade statistics. Factors such as the birth of the EU and the accompanying elimination of customs procedures among member states, the collapse of the Soviet Union and other socialist nations, and the forging of economic alliances and creation of free trade zones such as the EU and NAFTA, have not only affected trade relations among various nations, but also affected, in a number of ways, the circumstances surrounding the establishment of a trade data system. In addition, the Standard International Trade Classification (SITC), which forms the basis of world trade assessment, was revised for the third time in 1988. This third edition raises the new issue of how to maintain time series continuity in commodity trade statistics.

The Institute of Developing Economies (IDE), using AIDXT (Ajiken Indicators of Developing economies, eXtended Trade data), has been processing annual data from the twenty-four OECD nations, UN member states other than OECD members, and Taiwan (which belongs to neither) to create and utilize trade statistics matrices. All are based on data received from the UN, OECD and Taiwan in magnetic tape form, then processed at the IDE, which now maintains data by commodity and country (i.e., tariff zone) in time series format ranging from 1962 to 1992. In processing this data, the IDE has striven to adjust for such differences in each country's trade statistics as country or tariff zone of destination, monetary value, and quantitative unit in order to enable international comparisons and increase software efficiency.

However, it is now necessary to comprehensively re-examine current AIDXT data and the overall retrieval system in light of the aforementioned changes in circumstances surrounding world trade statistics. To do this, the IDE formed the Research Committee on World Trade Statistical Data and Retrieval Systems to examine, over a two-year period, the following three issues:

(1) Restrictions in the creation of trade statistics on the UN, OECD and Taiwan, and issues concerning matters arising from differences in these trade statistics;
(2) Issues concerning the problems and extrapolation processes involving database creation; and
(3) Issues concerning the use and application of trade statistics.

Below is a summary of each chapter.

Chapter 1

International Trade Statistics and Export Unit Price Deflator: towards a world trade model

Soshichi Kinoshita

This chapter reports an attempt to estimate sectoral export unit value index which can be used for modeling sectoral trade flows among countries within the multi-country model. First, we discuss briefly the general framework of trade flow model, quality of the trade data and the method of estimation. Then, we summarize our estimated results and make a preliminary evaluation on them.

The last two decades have seen an increasing interest in the economic interdependencies among countries or regions in terms of trade, capital flow, human resource and information. Reflecting this, there has been a rapid growth in the number of global multi-country models in recent years.

The pioneering grobal multi-country model was developed in 1970 under the Project LINK, which was launched by Prof. L.R.Klein and other leading econometricians in Europe, North America and Japan.

Formally a multi-country model is constructed to link together, in a consistent way, the national econometric models in each of the main countries or regions of the world. Here the idea of achieving consistency in the international linkage refers to the world trade accounting identities in the world market, which is briefly explained below.

Let T_{ij} be a trade from country i to country j, a trade share matrix, a_{ij} will be defined as

$$a_{ij} = T_{ij}/\sum_i T_{ij} = T_{ij}/M_j \ , \ M_j = \sum_i T_{ij}$$

where M_j is the total real imports of country j.

Thus, a_{ij} is the market share of country i's export

in the j-th country's imports. Then, by definition, the total real export of country i, E_i is given by

$$E_i = \sum_j T_{ij} = \sum_j a_{ij} \cdot M_j$$

Since the row sum of E_i is equal to the column sum of M_j by an accounting identity, world real exports EWT equal world real imports MWT as

$$EWT = \sum_i E_i = \sum\sum T_{ij} = \sum_j M_j = MWT$$

Turning from the real balances to the nominal ones, nominal value of the j-th country's imports, MV_j is the column sum of nominal trade flows, TV_{ij}, and is expressed as

$$MV_j = \sum_i TV_{ij} = \sum_i T_{ij} \cdot PE_i \ , \ MV_j = M_j \cdot PM_j$$

where PE_j is export prices and PM_i is import prices.

Combining the above two identities, PM_j is derived from PE_i as

$$PM_j = \sum_i a_{ij} \cdot PE_i$$

If a_{ij}, PE_i and M_j are given, E_i and PM_j are consistently transformed with the identity that world exports equal world imports in both real and nominal terms.

For the dynamic analysis of international linkage among countries or regions outlined above, a nominal trade flow matrix and export price index must be constructed using the common commodity and country classification. The basic data sources for these are the UN Commodity Trade Statistics and the OECD Foreign Trade Commodity Statistics. Both of these foreign trade statistics provide value and quantity data classified according to the SITC nomenclature by country.

In international trade statistics, a trade flow between any two countries is recorded twice; in the exporting country and importing country. As a result, bilateral export and import data can be and has been checked against each other. There have been pointed a number of discrepancies in reported data, which can be classified into three categories:

1) differences in valuation,
2) differences in reporting system and
3) errors.

The first one is related to the f.o.b. valuation of exports compared with the c.i.f. valuation of imports. Usually imports (c.i.f.) is recorded by 5 to 10% larger than the corresponding exports (f.o.b.). Additionally, discrepancies are caused by the time lag needed for transport and the difference in the exchange rate used to convert domestic currency to the US dollars.

The second category includes definition of partner countries and coverage of traded goods. As regard to the former, when trades between two countries are made througt entrepot in the third country, exporting country defines the third country as a trade partner, while importing country defines country of original producer as a partner. Sharp trade discrepancies between the US and China since around 1985 can be explained by the different definition of their trades with Hong Kong. While the US records imports of Chinese commodities through Hong Kong as imports from China, China records exports of domestic products to the US through Hong Kong as exports to Hong Kong. The same explanation is applicable to the case for trades between Japan and China.

The last factor is the errors in the recording and processing the data in individual countries. Imports and exports may be under-valued to save tax payment if high import or export duties are levied on commodities par value at the custom.

International trade statistics by UN and OECD are organized to cover both value and quantity data of classified commodities traded for the reporting countries. The current state is, however, that as for the agricultural and other primary commodities, both developed and developing countries report value and quantity data, while most of the developing countries report quantity data only for the selected manufacturing commodities. As a result, the use of unit value approach is very unsatisfactory for the manufacturing products to estimate the export price index of developing countries. This difficulty may be partially solved when the reporting countries in the developed area record both quantity and value data of exports from the corresponding developing countries. Anyway, the limited availability of quantity data reported to UN by developing countries is one of the major difficulties in using unit value approach to export price estimation.

Keeping in mind the shortcomings of the UN data, an attempt was made to estimate the export unit value index for 30 countries and 20 commodity groups listed below. The period coverd were 21 years from 1971 to 1991.

The countries covered include 23 OECD countries, two Asian NIES (Korea and Hong Kong) and five ASEAN countries (Indonesia, Malaysia, Philippines, Singapore and Thailand). Of the twenty commodity groups, 18 are for manufacturing products classified

in the two-digit level, and the remaining two are agricultural and mining industries.

Estimation of unit value index was made in the three steps: first, export unit values by commodity were computed over overlapping four-year subperiods using value and quantity series at the three to five digits commodity classification. Second, the detailed unit values were aggregated into 20 commodity groups based on the Laspeyres and the Paashe formulae. Lastly, unit value index for the subperiods were linked together to obtain a consistent time series with the common base year of 1980.

355 series of unit price index are successfully estimated, which is about 55% of 600 series, the total number of sectoral series to be estimated by country. As a whole, results of the ASEAN and Oceania countries are incomplete in the manufacturing sectors mostly due to the small shares and lack of sufficient quantity data. The number of disaggregated commodity items used in the sector aggregation varies among countries and commodity groups, ranging from 1 to 18. The larger number of items is used in the aggregation in the case of agricultural and six manufacturing products (food, textiles, chemicals, iron and steel, general and electrical machineries).

Quality of the estimated price series are examined in a different way. First, by comparing the Laspeyres series with the Paashe ones by commodity group, it is found that two series have a close correlation, shown by R^2 larger than 0.95 for 90.3% of the total series. Only 5% of them shows a correlation coefficient less than 0.9.

Second, the similarity and difference of sectoral price indicies across countries is checked with the correlation analysis between national series and world average ones. About 64% of the two series is highly correlated with each other showing R^2 in the order of 0.9, and the commodities with R^2 larger than 0.8 are about 84% of the totals.

Third, estimated export unit value indices are compared by sector with the corresponding price indices obtained from individual national sources such as producer price index and export price index. Elasticities of producer price or export price index with respect to the UN unit price index are computed for G7 countries, Netherlands and Korea. It is shown that the elasticities of the light industry products are concentrated around unity, while those of machinery products differ significantly from unity. The latter result implies that unit price index reflects not only demand-supply conditions but also quality changes in the product.

Chapter 2

The Use and Applicaiton of Foreign Trade Data at the United Nations

Itsuo Kawamura

This chapter does not deal with collection and publications of foreign trade statistics by United Nations system (on Chapter 3) but discuss how the UN Dept. of Economic and Social Information and Policy Analysis (UNDESIPA) makes use of these data in its research and analysis.

This chapter first discuss briefly nature of LINK Project, its central linkage mechanism and its basic trade data. Focus is on the methods and problems of preparing its foreign trade data (foreign trade matrix).

It also discuss a new direction in the construction of manufacturing export price indexes.

The LINK project, used and maintained by the UN Department of Economic and Social Information and Policy Analysis, comprises a set of large-scale national econometric models that can be made to interact so as to generate consistent forecasts of world production levels, commodity trade patterns, price relationships, and balance of payments positions.

The LINK system is based on the premise that teams of national modellers are in the best position to assess trends in this country and build an econometric model suitable for forcasting its direction. The principal linkages would come through merchandise trade flows and prices. Imports would be determined endogenously in each national model and exports would be allocated through a central trade matrix, with the world trade identity satisfied in the global solution.

80×80 trade matrices of the bilateral trade in SITC 0+1, SITC 2+4, SITC 3, SITC 5-9 between 70 individual countries and 10 regions are constructed at UNDESIPA. Total merchandise trade (0-9) is derived by summing the four matrices. These have been estimated yearly, starting 1965, generally with a

two to three year lag.

Construction is initiated by obtaining from UN Statistical Office (UNSO) trade matrices that are used to produce SPECIAL TABLE:C, World Trade by Commodity Class and Regions. These matrices show the reported trade by country and estimated by UNSO of 189 countries to 22 individual countries and regions for 21 SITC categories. The raw data are transferred into a binary mode matrix aggregated into 80 LINK countries and regions for the four SITC categories. Hence, the first stage of processing produces 80×22 matrices for each of the above SITC categories. At the second stage of processing, the 22 destinations are extended to the 80 LINK countries and regions by applying the previous years shares of countries in the aggregate grouping to the current year's grouping.

At the third stage of processing, the data for countries which have reported exports for the current year are extracted from the UNSO COMTRADE database and used to replace the estimates in the last stage of processing, with reported data for countries where it exists. The country totals for each of the groupings are compared to the totals given by the original UNSO file and adjustments are made to bring two totals within a 3 % difference. To do this use is made of the imports data in COMTRADE so that exports may be captured by partner country data. If partner country data do not exist, efforts are made to find other estimates from secondary sources such as reports of central banks, industry publications, country experts, specialized agencies, etc.

Recent efforts in the LINK foreign trade data requires the creation of a wider series involving annual imports data for all countries for a "movable" 15 year period, starting with 1976-1990.

The first step is the creation of a table of total imports trade for all countries for all years. Total imports value is extracted for all COMTRADE reporters. For country/periods not in COMTRADE, IMF/IFS figures are used, and, where these are not available, values are taken from UNSO Non-D files or from publications such as UNCTAD table for least diveloped countries.

The next step is to bring in any available external data such as UNSO energy import data (from Energy Yearbook), FAO trade data (from the FAO Trade Yearbook files) and CMEA data on the Trade of USSR.

One other sources of trade data in UNSO "Non-D tape" which contains data of essentially government origin but at insufficient detail for inclusion in COMTRADE.

Inverted COMTRADE is the source of most of the file for each non-reporter. In the LINK an import matrix is constructed and the inverted data is therefore exports of COMTRADE reporters. Inversions are also produced for the UNSO non-D data, the CMEA USSR data where these would be useful and any other data at the partner by commodity level. Where data is available for one year and not others "mirror image" data is produced on the expectation that similar trade would happen in the "missing" year.

The quality of a unit value depends on the stability of the product mix under its code. If that mix remains relatively constant, then the unit value can be a good proxy for the average price change in that category. Quality can also be affected by errors in the data themselves. For example, if a quality or a value is entered into COMTRADE with the decimal point in the wrong position, it will follow that the unit value will also be wrong. One way to mitigate this type of error is an numerical filtering. One problem in employing numerical filters is distinguishing between unit value fluctuations emanating from product mix changes and fluctuations derived from erroneous input. A second problem is turning the filter to screen properly. A too restrictive filter will yield a nicely behaved index but increases the risk of rejecting good data, while a filter which is too lax increase the risk of accepting bad data. The quality of the final index is highly correlated with the number of included in the export basket and the number of observations sampled for each items. It is preferable to use partner country data to construct export price indexes. These yield many more observations per 4-digit SITC code than reported exports. A test of the DESIPA filter for 17 sub-index of manufacturer was made for the Republic of Korea and Brazil on their reported exports by the two countries and on their exports derived from the reported imports of trading partners. For categories that had more than 20 observations, the index was constructed both with and without filters. If the number of observations was less than 20, the filter is turned off. A regression

equation is estimated between the corresponding indexes of both countries to check the correlation of the price movements under the null hypothesis that for each category, the price movements are influenced by world price. The procedure for using partner country data is to first download reported imports of the desired category at the 4-digit level from the country to all partner countries and then to add together all quantities and value under the same code. Hence there is only one (average) unit value for each 4-digit code. This would be passed through the filter and if it survives the filtering, it would be used for the computation of the indexes.

We derived a summary of the regression equation which were run between the corresponding indexes (partner country data). For SITC 5, the correlation is fairly high using Paasche index. The correlation is also strong for most of the indexes in SITC 6, which, like SITC 5, consists of relatively homogeneous goods, mainly manufactured goods classified by the material they are made from. In SITC 7, we see the least correlation. The basket of goods in any one country's 4-digit code in this category depends on its level of development. It strengthens the case for using country specific unit value indexes in this category.

In views of the need for consistent and timely availability of the foreign trade data of all countries, it would be advisable that efforts be made to provide technical assistance to upgrade statistical capability of developing countries.iIt is also desirable for various UN and other agencies and research institutions to co-operate and coordinate more efficiently in the creation of databases. Japan would increase its contribution of funds and personnel to the collection and estimates of various international statistics at multilateral institutions.

Chapter 3

The Trade Statistics of International Organizations

Hideki Hiraizumi

This chapter presents an overview of the collections of statistics on world commodity trade that have been released by major international organizations, and is intended to facilitate the utilization of these statistics collections.

In foreign trade statistics, an important aspect of economic statistics, the results (i.e., monetary amounts and quantities) of a nation's economic activities within a certain period (month, quarter or year) are calculated for each trading partner, item, and year, etc. Statistics concerning foreign trade encompass not only something called customs clearance statistics, but also statistics on export letters of credit received, export validation statistics, and import approval report statistics, for instance. Generally, the term foreign trade statistics refers to customs clearance statistics, which deals with the determination, at the actual point of customs clearance, of the quantity and monetary value of goods actually imported and exported.

Currently, the foreign trade statistics created by the majority of countries are customs clearance statistics, which are used by international organizations to prepare the aforementioned international trade statistics collections. In contrast, the foreign trade statistics generated by many of the former socialist states were based not on customs totals, but rather on reports submitted to ministries of foreign trade by authorized foreign trade organizations. Since converting to market economies and joining the world economy, however, these nations have been reverting back to customs clearance statistical methods.

At the same time, the abolishment of customs procedures within the EC area following market union on January 1, 1993, has forced EC nations to rely on data on value-added taxes paid by businesses on a monthly basis (instead of using customs clearance statistics) for statistics on trade within the EC.

What follows is a discussion of cooperation between international organizations and national statistical organizations. The primary trade publications of the United Nation Statistical Division are (1) "International Trade Statistics Yearbook" (Vol. 1, "Trade by Country"; and Vol. 2, "Trade by Commodity") and (2) "Commodity Trade Statistics," which contains annual data by reporting country.

Publication (1) above lists data by country (Vol. 1) and by commodity (Vol. 2). First published in 1952, this publication lists data reported by the government of each country. Figures reported in that nation's own currency or commodity classifications are converted

to U.S. dollars (using a predetermined rate) and to the classifications of SITC.

Publication (2) contains data reported by member nations on an annual basis to the United Nations Statistical Division. Data collected by the UN that is in a country's own currency is converted (using a standardized format) to facilitate international comparison. The predetermined rate used to convert figures in a country's own currency to U.S. dollars is listed in "Monthly Bulletin of Statistics," which is published by the United Nations Statistical Division, and which is more detailed than tables 4 and 5 in publication (1) above with respect to breakdowns by trading partner. SITC is the system of commodity classifications. There are also explanatory notes on each nation.

Below is an overview of the UN's regional economic and social commissions and the trade statistics collections published by each.

(1) The ESCAP covers the Asian-Pacific region and publishes "Trade Statistics of Asia and the Pacific - Yearbook." Data is provided by the UN Statistical Division, and the commodity classifications are those of SITC. Table 1 shows import and export totals by country of origin and country of destination; Table 2, exports and imports on a one-digit level of SITC·R2; Table 3, exports and imports on a three-digit level of SITC·R2; and Table 4 exports and imports according to the UN's Classification by Broad Economic Categories (BEC).

(2) The ESCWA, which covers Western Asia, publishes the "External Trade Bulletin of ESCWA Region - Yearbook," which contains data generally supplied by each nation's statistical organization, and from the UN's "Commodity Trade Statistics" and the IMF's "Supplement on Trade Statistics," for example. This publication is divided into three parts: part one, a condensed table on foreign trade in the ESCWA area; part two, transactions between ESCWA nations; and part three, the exports and imports of ESCWA member states, presented in accordance with the major economic national and regional classifications used by the UN Statistical Division.

(3) Covering all of Africa, the ECA publishes three different series of African foreign trade statistics: series (a), (b), and (c), which are respectively subtitled "Direction of Trade," "Trade by Commodity," and "Summary Table." All three are published irregularly. Series (a) features a breakdown by trading partner of export and import totals of the 37 ECA nations over the past 10 years. Series (b), which contains the same information as UN Statistical Division's annual "International Trade Statistics, Vol. 2," although to the five-digit level of SITC. Much of series (c) consists of time series of ECA nations' import and export totals to and from major nations and regions, and matrices of transactions within the ECA.

The UNCTAD regularly publishes (1) "UNCTAD Commodity Yearbook," and (2) "Handbook of International Trade and Development Statistics." Publication (1), published since 1984, was released under the title "Yearbook of International Commodity Statistics" in 1984 and 1985, and is characterized by statistics on worldwide, regional and national levels of the trade in, production of and consumption of major farm products, minerals, ores and metals.

Sources of data include the publications of national governments and organizations such as the UN Statistical Division, the IMF, and the FAO Statistical Division and Commodity Trade Division.

Sources used for publication (2), which contains statistics needed by governments, universities and researchers to analyze worldwide trede and development are the official publications of international organizations and governments.

"FAO Trade Yearbook," published regularly by the FAO, contains export and import total for farm products, along with quantities, unit prices and price indices for foods (part 1); trade in farm products (part-2; SITC commodity categories are used); trade in farming implements, fertilizer and other farming equipment and supplies (part 3); and agricultural trade value by country over the past 6 years (part 4). Part 4 deals with two-digit-level major farm products and farming equipment and supplies. Major sources for this publication include data received from and publications released by governments and the UN Statistical Division, and countries' replies to FAO questionnaires, although the FAO makes its own estimations and evaluations (based on data from a nation's trading partners) when reliable, up-to-date data is unavailable. Commodity classifications are those of SITC·R2. To preserve time-series continuity, data in SITC·R3 format is converted to SITC·R

2.Notes are included concerning the relationship of individual products under SITC·R2 and -R3.

The IMF releases a monthly, quarterly and yearly series titled "Direction of Trade Statistics." The yearly report features data not only on individual countries, but rather totals for different levels of economic development and different geographic and economic region. These classifications harmonize with the United Nations' national and regional classifications and with those of the IMF's International Financial Statistics (IFS). This publication deals only with national and regional export and import totals. As most data are in each country's own currency, the IMF converts these figures to U.S. dollars.

The OECD'S major publications are "Monthly Statistics of Foreign Trade" (series A), "Statistics of Foreign Trade" (series B), and "Foreign Trade by Commodities" (series C). Series C, which deals with OECD member states, contains detailed national regional and breakdowns of each country's SITC section (one digit) and division (two digits) over the past six years. OECD nations switched to SITC·R3 in 1988, and previously released data has been converted from the R2 format to R3.

Chapter 4

Commodity Classifications in Trade Statistics

Yasuko Yamamoto

One type of information needed to use trade statistics is knowledge about the commodity classifications used in trade statistics. Although knowledge about item classifications is unnecessary when using only export and import totals, knowledge of the commodity classifications used in trade statistics is needed to determine (from individual nations' trade statistics) and utilize trends in monetary values and quantities of commodities and groups of commodities imported and exported. Thus, this chapter is a guide to commodity classifications for persons who use international trade statistics.

The following three types of commodity classifications are used by nations in preparing trade statistics.

(1) A classification system based on the Standard International Trade Classification (SITC) of the United Nations.

(2) A classification system based on BTN, CCCN and HS, devised by the Customs Cooperation Council (CCC).

(3) Classification systems used by individual nations.

This chapter discusses types (1) and (2): their histories, characteristics, correlation, the effects of revisions in commodity classifications on the use of time-series trade statistics, and problems with continuity.

In section 1, "SITC," which discusses the history of SITC revision and presents an overview of the changes made, the author states that SITC began with League of Nations' "Minimum List of Commodities for International Trade Statistics." This subsequently became SITC Original, which was in turn revised to produce SITC Revised, SITC Revision 2, then finally SITC Revision 3.

Section 2, "BTN and CCCN," discusses the history of the Customs Cooperation Council and the process by which many of the world's nations have come to adopt the nomenclature of BTN and CCCN (internationally agreed-upon commodity classification tables for tariff rate tables) for their trade statistics classification tables.

Section 3, "HS," explains the HS standard that has been adopted by 119 (as of October 29, 1993) nations for their trade commodity classification tables. Commodity classification tables for trade statistics have been consistently prepared under guidance of the Customs Cooperation Council (CCC) since the organization's founding; the UN's SITC was drafted to conform with the CCC's commodity classification code. And while most of the world's nations (with the exception of the former Soviet Union, Eastern Europe and other socialist nations) had adopted commodity classifications that comply with CCC nomenclature, Canada and the U.S., which occupy important positions in world trade, had used their own trade commodity classification systems, which created various difficulties in tariff negotiations and trade friction. For this and other reasons, including the need for a new system of classification codes in response to technological progress and changes in the trade structure, a group of nations (primarily CCC members) reached an agreement regarding trade statistics for release in 1988 for beyond: namely, that beginning with countries where application was

possible, trade statistics would be converted to HS standards before release. HS, a major revision of the four-digit code system of CCCN, comprises 5,019 six-digit codes.

Section 4 is titled "The Characteristics of and Relation Between SITC and CCCN/HS." SITC, in which all commodities are classified under a basic heading with a five-digit (sometimes four-digit) code, makes it possible to classify products under a four-digit subgroup, three-digit group, two-digit division, and a one-digit section (between 0 and 9). The most important advantage of SITC is its structural conduciveness to economic analysis: SITC is designed so that all commodities, from raw materials to final products, have a classification code with a first digit that reflects the processing stage or industrial origin.

CCCN and HS, in contrast, were designed primarily for tariff rate table applications, and so raw materials and the products made from them are classified under the same section, or a section with a consecutive code number, thus preventing integration according to first digits, and making the system unsuited to economic analysis, etc.

"The Effects of Commodity Classification Revisions on the Use of Time-Series Trade Statistics and Problems with Continuity"(section 5) points out, using actual examples, the problems that changes in commodity classifications cause in the use of time-series trade statistics. First, the degrees of continuity between (1) SITC Original and SITC Revised, (2) SITC Revised and SITC Revision 2, and (3) SITC Revision 2 and SITC Revision 3 is explained (with emphasis on the major changes in the classification framework between SITC Revision 2 and SITC Revision 3) with examples and as they relate to the principles of classification. This discontinuity originates in the discontinuity between CCCN and HS: the former correlates with SITC Revision 2; the latter, with SITC Revision 3.

Lastly, the author expresses her thoughts on which type of commodity classification table is suited to commodity classifications for database trade statistics. Of course, it would be ideal for every country in the world to adopt the same standardized commodity classification table. While the country-by-country trade statistics now compiled by international organizations all comply with one version of SITC, the author states that HS data satisfies the needs of database users to a greater degree.

Chapter 5

Correlating and Converting Classifications of Disparate Systems: an attempt at commodity classification conversion using grouping and cutting

Yosuke Noda
Yasuko Yamamoto

As more and more countries switch to HS for trade commodity classification under multilateral agreements, more of the trade statistics data provided by international organizations are in the SITC·R3 format (which complies with HS), instead of SITC·R2. Consequently, methods of converting from SITC·R2 to SITC·R3 (or vice-versa) are needed to assure the continuity of time series data. The objective of this chapter is to consider a method for converting the statistical values of classification code A to those of classification code B (assuming correlation between codes A and B from different classification systems), and then apply this method to correlation between SITC·R2 and SITC·R3.

In order to relate the classifications of different systems, a correlation code table that indicates correlation between the two is needed. An important issue in the use of a correlation code table is determining what types of correlation exist between the two classifications. Here, the term "group" is used to indicate a collection of classification codes that have a closed correlation and which form the heart of classifications in the correlation code tables. Thus, of the two classifications presented in Chapter Seven, a group is a collection of classifications that correspond to the finest common derivative (FCD). Thus, a group begins at a common connection (there must be at least one) in the correlations and ends at the component element at which the connections cease. When making formal groups, group commonness can become a problem, but these common characteristics are ensured by cutting (discussed below) to create subgroups.

Correlating statistical values for each group eliminates the need for distributing the statistical

values of each classification code, thus preventing the estimation error that would result from distribution. Therefore, converting statistical values for each group is the method used in this chapter. In short, instead of thinking of correlation between classifications in terms of separate classification codes, a group, which is a closed collection of separate, related classification codes, is considered as a single correlation.

Although considering a group as a single correlation makes it possible to reveal the characteristics of groups with small numbers of elements, as the number of elements grows larger, the range of the group's characteristics could conceivably broaden and the characteristics of common elements weaken to the point that group content would become ambiguous. Consequently, when considering grouping, it is not only necessary to make group correlation clear, but also at the same time to subgroup into sufficiently small collections as a way to handle groups that are too large.

When a correlation is removed from a group and the remaining correlations are divided into several subgroups, the group is said to have been cut, and the correlation removed is referred to as the cut element. Subgrouping is a means of regrouping a correlation code table from which several correlations have been cut.

Cutting determines the characteristics of the correlation code table, and so subgrouping via cutting can be considered as a model for correlation code tables. Actual applications of this method to be considered in this chapter are GRT_{32}, a basic model for commodity classifications (in which no cutting is performed), and GRT_{32}(IDE) an original model with cutting created by the IDE. Although many methods of cutting are conceivable, the method used in this chapter is one that involves separately examining each correlation in the group and readjusting those judged to be unrelated.

The correlation code table used for commodity classifications SITC·R3 and SITC·R2 is based on the correlations taken from the UN Statistical Division's "Standard International Trade Classification Revision 3" (Statistical Paper Series M No. 34/Rev 3, United Nations 1986). Basic model GRT_{32} is based on this correlation code table, which is comprised of uncut correlations.

The IDE's model, GRT_{32}(IDE), was created by cutting basic model GRT_{32} in accordance with the principles described below. The correlations of GRT_{32}(IDE) are shown in "GRT_{32}(IDE) and Table of IDE Standardized Country Codes." In creating GRT_{32}(IDE), correlations are typed for cutting according to individual commodity characteristics in order to clarify the specific method of cutting. In the correlation code table for SITC·R2 and SITC·R3, the revised and newly added classification standards discussed in section 4 are intricately intertwined. When considering the individual elements that make up a group, this intricately intertwined relationship almost always results in connections of varying levels of strength between SITC·R3 and SITC·R2. Relationships that affect this strength include the following.

(1) Complete agreement in the range of the commodities included;
(2) Partial correspondence between individual commodities included in a one-item code (i.e., varying rates of correspondence);
(3) Commodities that are from the same raw material, or whose raw materials are significantly or partially related;
(4) Related functions or applications;
(5) Related manufacturing or processing;
(6) Future correlation between commodities is expected, although currently there appears to be no correlation between specific commodities.

Of these, type (1) is viewed as an assuredly strong connection; types (2) through (5) as anywhere from strong to very weak, depending on the elements in question; and type (6) as weak.

Elements falling under type (6) are the first candidates for subgroup cutting, followed by weakly connected elements in relationship types (2) through (5). Only those candidate elements that qualify as subgrouping factors are cut. In determining cutting candidates, emphasis is placed on subgrouping into groups of discernible characteristics.

Conversely, if group characteristics are clear despite an extremely large number of group component elements, even elements that are weakly connected are not candidates for cutting (as long as they do not disrupt group characteristics).

Some groups can have extremely large numbers of elements, in which case extreme cutting is necessary

to make group characteristics more discernible.

This even made it possible (to a certain extent) to subgroup fibrous thread and textiles according to raw material, and to integrate clothing, plastic products and rubber products into their own respective groups. Nevertheless, the standards for cutting are not always the same for each commodity.

The experimental cutting performed for this chapter is still in the initial stages; several issues remain to be addressed. These are:

(1) Considering the standards for cutting and readjusting the elements to cut.

(2) Some products that would customarily belong to a single group happen to completely correspond in a separate group somewhere between SITC·R2 and SITC·R3. In such cases, the integration of such groups is required to maintain as much continuity as possible between product groups in the same range.

(3) For correspondence tables for SITC·R2 and SITC·R1, the aggregate relationship between three-, two-, and one-digit codes is maintained, but this problems has not been solved for GRT_{32}(IDE). Thus, the restructuring of groups (including hierarchical structures) is necessary to create tools for applications requiring the aggregate use of trade statistics.

(4) The ultimate objective of grouping is to complete some sort of guide to the use of the IDE's standardized commodity classifications, which would systematize, in accordance with common concepts, the commodity classifications of two different systems.

Chapter 6

The Division and Unification of Nations: applying temporal data models

Hidekiyo Sakamoto

In this chapter we will consider the expression and use of knowledge concerning the unification and division of a country or custom area (hereinafter simply "country"), which is a type of AIDXT metadata. Trade statistics track the flow of commodities from one country to multiple trading partners over a certain period, and so information on countries is important in the use of statistics.

As the AIDXT database consists of UN, OECD, and Taiwan trade statistics, the country categories used by one statistical organization do not always correspond to those of another. Moreover, with the passage of time, countries undergo unification, division, and changes in name and political structure. For instance, while the point at which a reporting country or trading partner is listed in trade statistics as having unified or divided may differ from the actual time of unification or division. Thus, when using trade statistics to process data, it is necessary to adjust country categories and possess knowledge of countries' unification and division and of the actual points in time of unification or division. To consider changes in country codes it is necessary to standardize country code categories.

The IDE Standardized Country Codes are a system of country classifications created to standardize the country code systems used by the UN and the OECD.

The IDE Standardized Country Code Table (Table 2-1 in "Table of IDE Standardized Country Codes") permits comparisons of UN' and OECD' country categories, but does not incorporate changes over time in countries. Because of historical changes in the sphere of a country (e.g., birth, division and unification), changes over time in country categories must be considered if trade statistics are to be used as a time series. Table of Changes in IDE Standardized Country Codes (Table 2-2 in "Table of Changes in IDE Standardized Country Codes") reflects the changes in countries over time.

As definitions of partner countries are left to reporting country's judgment, and because the points at which a partner country is considered to have unified or divided are not always the same in each reporting country. Table 2-2 cannot provide accurate information for each reporting country. To do this, a number of Table 2-2 equivalent to the number of reporting countries would be required, resulting in a massive volume of data.

As a country's division or unification is related to time, "time" must be incorporated into data models in order to express this knowledge. A model that concerns time in this manner is called a time data model. The expanded data model presented in this chapter was developed by conceptually expanding the logical time data model of Segev and Shoshani, which was developed as a framework for data models designed to handle temporal data.

The basic concept of Segev's data model is the time sequence (TS). Each object has attributes of which values change over time, and TS is the term for the time-related ordered sequence of these attribute values, expressed thus:

$$TS = \{ x_s(t_1)\ x_s(t_2) \cdots x_s(t_n) \}$$

where $t_1 \cdots t_n$ represents time.

A class is a collection of objects with the same attributes, and a collection of TS's that belong to the same class is called a time sequence collection (TSC). In other words, a TSC is a collection of TS's represented thus:

$$TSC = \{TS_1\ TS_2 \cdots TS_n\}$$

	t_1	t_2	⋯	t_n
TS_1	$x_{s1}(t_1)$	$x_{s1}(t_2)$	⋯	$x_{s1}(t_n)$
TS_2	.	.	⋯	.
⋮				
TS_i	$x_{si}(t_1)$	$x_{si}(t_2)$	⋯	$x_{si}(t_n)$
⋮				
TS_m	$x_{sm}(t_1)$	$x_{sm}(t_2)$	⋯	$x_{sm}(t_n)$

where TS_1, TS_2, $\cdots$ TS_n are the TS's corresponding to such country objects as Japan, France and the U.S. and so on.

Segev's TS describes the changes of attribute values of a object over time, which can deal with country name changes. But it cannot represent country division or unification because of the necessity of dealing with relationships among multiple objects.

TSx and TSCx were developed as conceptual expansion of TS and TSC in order to handle country division and unification. A related object group which a TSx corresponds to is a collection of interrelated objects that unify and divide as time passes. A TSCx is a collection of TSx just as a collection of TS's is TSC. The basic operators of the expanded model produce new TSCx's from existing TSCx's. Repeating this process using the basic operators make it possible to perform complex operations.

Country division/unification and name changes can be defined thus:

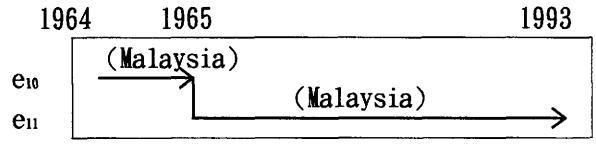
(1) A country name change is the change, at a certain point, of the country name attribute of an object in a country class.

(2) Country unification is the consolidation, at a certain point, of multiple objects in a country class into a new, single object.

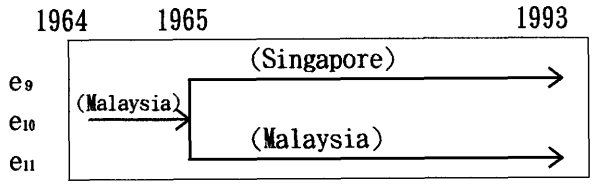
(3) Country division is the division, at a certain point, of a single country object in country class into multiple country objects.

(4) When part of a country joins with another country (or part of another country), this can be seen as a combination of division and unification.

In the case of Malaysia, for example, Malaysia devided into two countries, Malaysia and Singapore, in 1965. Using basic operators to get the information on objects of country name "Malaysia" from country classification TSCx for the years 1964 to 1993 produces the following results.



This shows that since 1965 Malaysia has been a different object. Using basic calculations to solve the object of country name "Malaysia" from country classification.



This shows that in 1965 object e_{10} divided into object e_9 and e_{11}, the respective country names of which are Singapore and Malaysia.

Chapter 7

Fundamental Cencepts of Summary Data and Inferences in Their Databases: using a world trade database as an example

Hideto Sato

This paper treats fundamental notions practicable for database systems that deal with summary data concerning social or regional information. Summary data here is data that is produced from atomic data by applying a certain summarizing operation. For example, take such data as "average wages by occupation." This data is summary data produced from atomic data such as payments of wages to invididual employees. Such summary data is often

called statistics, which is a major component of databases that deal with huge domains, such as objects in a whole country or events that occurred over an extended period of time. They are required not only by governments and international organizations but also by research and planning sections in enterprises and universities.

When dealing with summary data, if the corresponding atomic data are available it is possible to construct a database management system that presents to its users arbitrary summary data produced from the atomic data. This condition, however, is seldom satisfied in actual practice. Atomic data is often not available when a user wants its summary. In the above example, individual records on payments to employees are usually not preserved for a long time, for a physical or economic reasons.

This paper focuses on a database of summary data in such a situation.

A database is a collection of data shared by many users. Even summary data is rather easy to handle in a database, if there is consensus among users about the classification of objects related to the data. But if desirable classifications differ from one user to another,or if desirable classifications change with time, serious problems may arise. Heterogeneity in classification between a collector and an interrogator makes it difficult to derive the required data from the collected data. We call this difficulty the problem of derivability. Heterogeneity in classification among different collectors prevents direct comparison among the collected data. We call this difficulty the problem of comparability.

In preparation for treating these problems in an operational mannar, we first define a formal notion of summary data, by introducing two types of abstraction, categorization and summarization, on the basis of the set theory. An important point revealed by this formalism is that summary data may be derivable from another summary data by resummarizing the latter when the summarizing oparator is associative. This property allows summary data to be shared by many persons who have different points of view from one to another. In fact, widely shared summary data in the actual world are mostly described in forms of average, total, maximum, and minimum values, which are obtained by means of the respective associative summarizing operators.

Hence we concentrate on summary data produced by associative summarizing operators.

Next, we introduce the notion of a classification hierarchy in order to denote the above heterogeneity in classification. Then we propose three inference mechanisms applied to classification hierarchies. The first inference mechanism judges derivability of a classification from another classification and computes the reclassification rule between the classifications if the former classification is derivable from the latter. The second inference mechanism computes if the former classification is derivable from the latter. The second inference mechanism computes a special classfication - the strong finest common derivative (SFCD) - for a given pair of classifications. The third inference mechanisms compounds reclassification rules in different classification hierarchies for one reclassification rule in the compound classification hierarchy. Then we demonstrate how these mechanisms resolve the problems above.

As for the problem of derivability, we present that the judgement of derivability and that the derivation of required data from collected data can be carried out by the aid of the first inference mechanism if it is derivable. We also show how this mechanism improves a database of summary data in its usability and logical data independence.

As for the problem of comparability, we propose a kind of join peculiar to summary data on the basis of the SFCD computed by the second inference mechanism. Then we show how this type of join makes it easy to compare summary data related to heterogeneous classifications.

The third inference mechanism enables us to treat multi-classified (n-tuple) summary data as if it were binary summary data. This treatment makes it possible to apply the first and the second mechanisms to general multi-classified summary data.

These mechanisms are constructed on the basis of the set theoretical lemmata proved in the original paper. They need only two types of operations; finding a path in a hierarchy and composing relations. These operations are common to hierarchical systems and relational operations. Hence they can be easily implemented in a database management system for summary data.

(Note)

This paper is a summary of a thesis that has the

same title of this paper. The original thesis is written in 1982 by the auther. By this thesis, the auther was granted the degree of Doctor of Engineering from Tokyo University in 1983. In addition, the parts of this thesis were published as follows:

[1] Derivability and Comparability among Non-Atomic Data, in P. S. Glaeser (ed.) Data for Science and Technology, Pergamon Press, 1981 pp.565-568.

[2] Handling Summary Information in a Database: in Y. E. Lien (ed.) Proceedings of the ACM-SIGMOD International Conference on Management of Data, Association of Computing Machinery, 1981, pp.98-107.

In addition, the concepts introduced in this paper were enhanced in a theoretical manner in the following paper. For the recent development about the topic, please consult this paper.

[3] F. M. Malvestuto, A Universal-Scheme Approach to Statistical Databases Containing Homogeneous Summary Tables, ACM Transactions on Database Systems, Vol.18 No.4, Association of Computing Machinery, 1993, PP.678-708.