Bakgrundsfakta

Statistiska centralbyrån Statistics Sweden

Improvement of the reliability in the estimated distribution keys in Intrastat

2007:3

Ekonomisk statistik

The series Background facts presents background material for statistics produced by the Department of Economic Statistics at Statistics Sweden. Product descriptions, methodology reports and various statistic compilations are exampels of background material that give an overview and facilitate the use of statistics.

Publications in the series
Background facts on Economic Statistics

2001:1 Offentlig och privat verksamhet – statistik om anordnare av välfärdstjänster 1995, 1997 och 1999
2002:1 Forskar kvinnor mer än män? Resultat från en arbetstidsundersökning riktad till forskande och undervisande personal vid universitet och högskolor år 2000
2002:2 Forskning och utveckling (FoU) i företag med färre än 50 anställda år 2000
2002:3 Företagsenheten i den ekonomiska statistiken
2002:4 Statistik om privatiseringen av välfärdstjänster 1995–2001. En sammanställning från SCB:s statistikkällor
2003:1 Effekter av minskad detaljeringsgrad i varunomenklaturen i Intrastat – från KN8 till KN6
2003:2 Consequences of reduced grade in detail in the nomenclature in Intrastat – from CN8 to CN6
2003:3 SAMU. The system for co-ordination of frame populations and samples from the Business Register at Statistics Sweden
2003:4 Projekt med anknytning till projektet "Statistik om den nya ekonomin". En kartläggning av utvecklingsprojekt och uppdrag
2003:5 Development of Alternative Methods to Produce Early Estimates of the Swedish Foreign Trade Statistics
2003:6 Övergång från SNI 92 till SNI 2002: Underlag för att bedöma effekter av tidsseriebrott
2003:7 Sveriges industriproduktionsindex 1913–2002 – Tidsserieanalys
The Swedish Industrial Production Index 1913–2002 – Time Series Analysis
2003:8 Cross-country comparison of prices for durable consumer goods: Pilot study – washing machines
2003:9 Monthly leading indicators using the leading information in the monthly Business Tendency Survey
2003:10 Privat drift av offentligt finansierade välfärdstjänster. En sammanställning av statistik
2003:11 Säsongrensning av Nationalräkenskaperna – Översikt
2003:12 En tillämpning av TRAMO/SEATS: Den svenska utrikeshandeln 1914–2003
2003:13 A note on improving imputations using time series forecasts
2003:14 Definitions of goods and services in external trade statistics
2004:1 Hjälpverksamhet. Avrapportering av projektet Systematisk hantering av hjälpverksamhet
2004:2 Report from the Swedish Task Force on Time Series Analysis
2004:3 Minskad detaljeringsgrad i Sveriges officiella utrikeshandelsstatistik
2004:4 Finansiellt sparande i den svenska ekonomin. Utredning av skillnaderna i finansiellt sparande Nationalräkenskaper, NR – Finansräkenskaper, FiR
Bakgrund – jämförelser – analys

Continued on inside of the back cover!

These publications and others can be ordered from:
Statistics Sweden, Publication Services, SE 701 89 ÖREBRO, Sweden
phone +46 19 17 68 00 or fax +46 19 17 64 44.

You can also purchase our publications at our Statistics Shop:
Karlavägen 100, Stockholm, Sweden

Background Facts

Improvement of the reliability in the estimated distribution keys in Intrastat

Economic Statistics 2007:3

Statistics Sweden
2007

Background Facts

Economic Statistics 2007:3

Improvement of the reliability in the estimated distribution keys in Intrastat

Statistics Sweden
2007

Previous publication	Published since 2001
Producer	Statistics Sweden, Business and Labour Market Department Box 24 300 SE-104 51 Stockholm
Inquiries	Frank Weideskog, +46 8 506 946 89 frank.weideskog@scb.se

ISSN 1654-0719 (Online)
ISSN 1650-9447 (Print)

ISBN 978-91-618-1374-2

URN:NBN:SE:SCB-2007-X100ST0703_pdf

Printed in Sweden
SCB-tryck, Örebro 2007:3

Preface

The burden on data providers is being reduced more and more in the foreign trade statistics of commodities within the European Union (Intrastat) and a greater share of trade needs to be estimated. This, in turn, creates the need for more comprehensive and improved estimation methods, which can meet the need for continued accuracy in statistics on foreign trade in goods. This report describes methodology work that has been carried out with the primary aim of improving the quality of estimated trade in the Swedish Intrastat system.

The work has been conducted during 2006 by a project group within the Foreign Trade and Industrial Indicators Unit at Statistics Sweden.

Statistics Sweden, March 2007

Lars Melin

Anita Ullberg

Contents

Summary

This methodology project has highlighted a range of improvement measures to raise the quality of the estimated Intrastat data. The work has primarily focused on identifying more accurate estimations in connection with the division of estimated trade into commodity groups and countries.

Companies that are obligated to provide data in Intrastat but that are non-response are considered as "enterprises with history" if data from previous months are available. Enterprises under the threshold value or enterprises that are obligated to provide data but that, in principle, never submit Intrastat data, are considered as "enterprises without history". We have here analysed the effect of a change in the criteria for classifying "enterprises with history", from enterprises which have at least one historical month available during the past six months to those with at least one historical month during the past twelve months. The results show, among other things, that around 100 more enterprises would be able to be estimated with history. The share of the value of non-response enterprises that is estimated with history increases then from 63 percent to 76 percent. Approximately 500 commodity codes, which were previously dispersed among different estimation groups for "enterprises without history", disappear with the implementation of the new criterion. The share of commodity codes in arrivals that needed to be revised by at least 50 percent decreases from 12.6 percent to 5.5 percent. In dispatches, this share decreases from 14.1 percent to 2.9 percent. The number of commodity codes in arrivals + dispatches that were revised by at least 50 percent decreases in total from 1 200 to 400 with the new criterion for history.

The size categories used for enterprises without history have been changed to 1: 0-4 SEKm, 2 :4-40 SEKm, 3: >40 SEKm. These were previously 1: 0-10 SEKm, 2: 10-100 SEKm, 3: >100 SEKm. When the non-response application was created in 1999, when the size categories were also determined, the conditions were different from those today. Non-response was considerably greater and the threshold values have been significantly raised since then. More of the very large enterprises who were obligated to provide information were non-response and were placed in the third size category. In addition, there were serious problems in the beginning to find out the addresses of the foreign data providers. Intense non-response work is carried out today focusing primarily on the largest and most significant enterprises, resulting in a very low level of non-response in the largest group.

A routine for the reclassification of enterprises that lack an industry classification code based on actual Intrastat data has been implemented. This has resulted in the reclassification of close to 350 enterprises based on actual trade. A very noticeable effect on value could be seen for the size categories with the largest dispatches enterprises, where 25 enterprises could be reclassified to a value of SEK 11 billion.

The groups used for enterprises without history has been studied. The groups are today largely created manually and a great deal of resources are required for this. Furthermore, a great deal of subjective judgements are

made when classifying the groups. The new classification method is automated and the intention is that the groups should, in the long-term, be updated every month when the figures are produced instead of every three years, as is currently the case. The method used when creating the groups is cluster analysis. The manual process focuses almost completely on the industry codes included in the largest commodity groups while the cluster analysis method takes into consideration all the commodity groups.

Extra output controls for the estimated trade in commodity groups and countries are not currently carried out but should be implemented in the long-term. There are some ideas to implement controls on and continuous follow-up of larger revisions. Another proposal is to implement some kind of control of the ratios between estimated and reported values per country and chapter on a total level, or divided into estimated trade under the threshold value or with or without history.

1 Introduction

1.1 Background

The Intra trade relates to collected and estimated Intrastat trade. Estimated trade can be divided into estimated trade for non-response and estimated trade below the assimilation thresholds.

Non-response PSIs (providers of statistical information) and smaller companies that are under the thresholds are primarily estimated as total company values. In the next step, their estimated trade will be distributed among certain commodities and country levels. This division is made using a "distribution key", which is preferably determined from a company's previous reporting (for a company with history) and is otherwise decided from the reporting from similar companies (for a company without history). Companies below the threshold value are estimated using VAT information and their trade is broken down into commodity codes and countries, on the basis of similar companies. At the current time, it is not plausible to have distributed values on commodities and countries for each single company, since companies are grouped together when historical data is missing.

1.2 Objectives

The overall objective of this project is to evaluate and improve the reliability in the distribution keys. It also aims to develop methods that make it possible to get total figures (collected and estimated) on commodity and country levels for each company. A further objective of this project is to develop new methods in order to improve production and expand the dissemination and deliveries of data at company level. The implementation of new improved methods for the distribution keys of the estimated Intra trade makes it possible to obtain more information about collected and estimated trade on commodity and country level for each company.

1.3 Human resources used

The project began in June 2006 and was completed in February 2007. The work was carried out by Mr. Frank Weideskog (project leader) at the Swedish Foreign Trade and Industrial Indicators Unit (UI) and Ms Tiina Orusild at the Methods Unit, both in Business and Labour Market Department (NA). Additional people with expertise in the fields of methodology, IT and Intrastat production were also involved.

1.4 Description of the operation

An introductory administrative briefing meeting was held, followed by four project group meetings. During the project, four control meetings were held between the project leader and the customer.

The following main tasks have been carried out in the project:

- Carry out evaluations of the present distribution keys, and criteria for these, used in the Intrastat estimation system.
- Develop new methods for grouping companies together automatically based on similarity.
- Develop new methods to make it possible to provide total values for each company (collected and estimated) on commodity and country levels.
- Investigate if any seasonal adjustments should be made when dividing the estimated trade.
- Carry out improvements of the output checking of estimated trade divided by commodity.
- Implement the necessary changes in the Intrastat estimation system.
- Implement new routines for follow-up of divided estimated trade.

2 The Swedish Intrastat system

2.1 General description

The Intrastat survey is a monthly collection of arrivals and dispatches of goods within the European Union (EU). It takes the form of a 'cut-off' survey, in which 97 percent of the total trade within the EU should be included and the rest should be estimated. From 2005, Sweden has a threshold value of SEK 2.2 million for arrivals and SEK 4.5 million for dispatches. Coverage in the Swedish Intrastat system in 2005 was 97.4 percent for arrivals and 97.9 percent for dispatches. The Intrastat threshold is defined as a continuous twelve-monthly value based on VAT data. A company that is not required to report information can only be identified as such after delivery of VAT details from the Swedish National Tax Board. The company will then be notified of its obligation to report and will receive information on what this involves.

The Intrastat data to be submitted are the member state of the arrivals and the member state of the dispatches, the nature of the transaction, the statistical commodity code, other quantity, net mass and invoice value. It should be noted that other quantity data are not required for some commodity codes and for others, data on mass do not need to be submitted. Information can be submitted on a paper form, or via electronic media such as IDEP (Intrastat Data Entry Package). About 30 percent of the providers of statistical information (PSIs) currently report Intrastat electronically.

The last day for reporting is 10 working days after the end of the accounting period (month), in accordance with a specific timetable. Statistics Sweden must therefore have received the form no later than 10 working days after the end of the accounting month in question. Companies who report via IDEP have however one or two extra days to submit their information.

Intrastat data are expected from about 15 600 PSIs, or 12 400 individual companies, which are together responsible for more than 360 000 commodity items. Every year, more than 60 000 VAT-registered Swedish companies make some form of goods transaction with another EU Member State. About 21 000 of these have regular EU trade every month. VAT data are supplied to Statistics Sweden from the National Tax Board once a week.

Aggregated Intradata are delivered to Eurostat about 25 days after the reporting month and detailed Intradata 55 days.

2.2 Brief description of the Intrastat estimations in SE

In this section, the imputation methods used to estimate non-response and trade below the threshold value are presented.

For non-response companies, the total arrivals and/or dispatches of goods are first estimated using five different methods. One of these methods is then chosen and the estimated trade according to the chosen method is divided into commodity codes (CN8) and country. The division into commodity code and country is done with a distribution key, which is first decided by what the company has reported previously and then according to similar companies. Companies below the threshold value are estimated using VAT information and their trade is broken down into commodity codes and countries, on the basis of similar companies. When VAT information is missing a method is used where average trade value is calculated (see method (v) later in this section).

Arrivals and dispatches of goods are treated separately, which means that estimations of total arrivals and dispatches and the distribution keys are also carried out separately. This is not always clear from the notation that is used.

Estimation of total arrivals or dispatches

Let y_{im} be the total arrivals or dispatches for company *i*, month *m*. If company *i* is part of the non-response, the total arrivals and dispatches of goods are estimated by the methods *(i)* - *(v)* below.

(i) Exponential smoothing

y_{im} is estimated by

$$\hat{y}_{fim} = \alpha \cdot y_{i(m-1)} + (1-\alpha) \cdot \hat{y}_{fi(m-1)} \tag{1}$$

where $y_{i(m-1)}$ is the reported total arrivals or dispatches of goods for

company *i*, month *m-1*, $\hat{y}_{fi(m-1)}$ is the estimated total arrivals or dispatches of goods according to (1) for company *i* , month *m-1*, and α is a value between 0 and 1, $0 < \alpha < 1$, $\alpha = 0.2$ is used in the non-response programme, which is the default value in SAS.

(ii) Extrapolation with seasonality

y_{im} is estimated by

$$\hat{y}_{f_s im} = \hat{y}_{fim} \cdot \hat{s}_m$$

where $\hat{y}_{fim}$ is the estimated total arrivals or dispatches of goods according

to (1) for company *i* for current month *m* and $\hat{s}_m$ is the seasonal component for month m. Seasonal components are calculated on industry level (the three-digit NACE).

(iii) Estimation using VAT value

y_{im} is estimated by

$$\hat{y}_{(VAT)im} = x_{im}$$

where x_{im} is the VAT value for company *i*, arrivals or dispatches, current month *m*.

(iv) Manual imputations
Manual imputations are preferably chosen over methods *(i)* - *(iii).*

(v) Average trade value with seasonal component

y_{im} is estimated by

$\hat{y}_{(a)im} = \bar{z}_{im} \cdot \hat{s}_m$ where

$\bar{z}_{im} = \frac{1}{n} \sum_{t=m-12}^{m-1} z_{it}$ is the company’s average trade value the last 12

months, $\hat{s}_m$ is the seasonal component for month *m* as in (*ii)* and *n* is the number of non-missing obervations for company *i* during the previous twelve months. The sum in $\bar{z}_{im}$ is over non-missing monthly values for company *i*.If there is less than three monthly company values the previous six months then $\bar{z}_{im}$ is set to 0. The method is based on total company values.

Selection of estimation method
Once the three estimates are produced, a selection is made of which method shall be used. As mentioned, manual imputations are preferred. If one of the methods *(i), (ii) or (iii*) are selected but the VAT value and the estimations from the forecasting models are missing, then method *(v)* below is selected.

The final stage in the imputation is to divide the companies’ estimated total arrivals and/or dispatches into commodity groups and countries. This division is made using a "distribution key", which is preferably determined from a company's previous reporting (for a company with history) and is otherwise decided from the reporting from similar companies (for a company without history). Companies below the threshold value are estimated using VAT information or with method *(v)* and their trade is broken down into commodity codes and countries, on the basis of similar companies. More details about the distribution keys are given in section 3.1.

3 Project activities

3.1 Mapping of present distribution keys

Once a company's trade has been estimated on a total level, their estimated total arrivals and/or dispatches are divided into commodity codes and countries. The division into commodity code and countries is done using "distribution keys" that are firstly determined by how a company has reported previously (company with history) and are then decided by how similar companies have reported (company without history). Companies under the threshold value are estimated using VAT data or average trade value (see 2.2.2) and their trade is divided into commodity codes and countries using a distribution key that is calculated using similar companies.

1. Division of estimated trade for companies with history:
Division according to history is based on the past six months' trade. The prerequisite is that there must be at least one reported month. A maximum of three of the six months are used to calculate the distribution key. The months chosen are the three most recent months. A distribution key is determined for every company, where every value in the key is calculated as a ratio of the value of the combination commodity and country and the value of all trade. Using these calculated shares, the company's estimated total monthly values are divided into commodity codes and countries. Arrivals and dispatches are treated separately.

Example
Assume that company i has reported values for month *m-3* to month *m-1* and let $y_{ij(m-u)}$ be the value of trade that company *i* has had during month *m-u* with commodity/country combination *j*. The share of trade to be divided by commodity/country combination *j*, p_{ijm} is estimated by

$$\hat{p}_{ijm} = \frac{\sum_{u=1}^{3} y_{ij(m-u)}}{\sum_{u=1}^{3}\sum_{j=1}^{J} y_{ij(m-u)}}$$

The value of trade for company *i* with commodity/country combination *j* during month *m*, y_{ijm} is estimated by

$$\hat{y}_{ijm} = \hat{p}_{ijm}\,\hat{y}_{im}$$

where $\hat{y}_{im}$ is the estimated total arrivals or dispatches for company *i* for the current month.

If the company has reported only two months previously, the shares are calculated according to these months. The distribution key is therefore determined by the months that are available using the criteria for the number of months as described above.

2. Division of estimated trade for companies without history
If there is no history, the total estimated arrivals and/or dispatches are divided according to the trade of similar companies. The companies are divided into groups and a distribution key for every group is produced. The groups are made up of companies that trade in the same commodities. The groups are formed by size (according to the annual VAT values) and the companies' industry. All companies are divided into groups. Trade from non-response companies (lacking history) and companies under the threshold value are divided into commodities and countries according to the group's distribution key. The distribution key of a group is determined by the responding companies in a group.

The groups are created as follows. Companies are divided into three size categories (based on the annual VAT values) for arrivals and three for dispatches (<10 million SEK, 10-100 million SEK, > 100 million SEK). The grouping is then as follows:

The groups are created as follows. Companies are divided into three size categories (based on the annual VAT values) for arrivals and three for dispatches (<SEK 10 million, SEK 10-100 million, >SEK 100 million). The grouping is then as follows:

1) The companies is divided into industries according to a three-digit NACE, hereafter called NACE3I, and their trade is grouped by NACE3 hereafter called NACE3T.
2) The industry that the companies belong to according to the NACE classification (NACE3I) is compared to the actual trade (NACE3T).
3) The NACE3T-groups with the highest value are considered to be representative for the industry and are compared with other industries (NACE3I) to find similar trading patterns.
4) The groups are created manually by combining industries with similar trading patterns. Every group should consist of at least five companies, and none of the companies in the groups should account for more than 50 percent of the trade. If one company accounts for more than 50 percent of the group's total trade, this trade should not be included in the group. The industries containing less than five companies or industries that cannot be combined with other industries are placed in a miscellaneous group.

For each group of responding companies, a distribution key is determined with shares per commodity combined with country. Every value in each key is calculated as a ratio between two sums for the company group, the value of commodity combined with country as the numerator and the value of all trade as the denominator. Arrivals and dispatches are treated separately but the calculations are carried out in the same way.

A group of responding companies is denoted by *g* and has a common key f_{gj} as shown below, where the sum over *i* refers to companies that belong to the group (group affiliation is shown by $\in$)

$$f_{gj} = \frac{\sum_{i \in g} y_{ij}}{\sum_{i \in g} \sum_{j=1}^{J} y_{ij}}$$

3.2 Changes in criteria

In this section, we will study the effect of changing the two criteria when the estimated trade of companies with no history are divided into commodities and countries. The change of criteria resulting in these being considered as a company with history or in a change of size category.

We will first study the estimated trade (non-response + trade under threshold value) for 2005 at the first publishing date for each month. Table 1 shows the division of the estimated trade value where historical data have been used and trade where historical data do not exist.

Table 1
Estimated trade divided into trade with or without historical data 2005

	Estimated with historical data (SEK billion)	Estimated without historical data (SEK billion)
Arrivals	13.2 (37 %)	22.5 (63 %)
Dispatches	8.1 (33 %)	16.1 (67 %)

On the first publishing date in 2005 (t+60 days), the estimated company totals are divided approximately up to 35 percent with historical data and 65 percent without historical data. It should be noted that the first publishing date was 10 days earlier from 2006 onwards (t+50 days) and that the relationship between trade divided with history and trade divided without history changed. Data from the first half of 2006 show that the estimated trade based on historical data accounted for 45 percent and trade without historical data for 55 percent of the total.

In principle, estimated trade divided by historical data refers to trade submitted by data providers who have submitted data fairly regularly while estimated trade without historical data refers to data providers who have rarely or never submitted Intrastat data, or companies who are under the threshold value and who are not obligated to provide data. Table 2 illustrates the divided trade without historical data, by non-response value or value for companies that are not obligated to provide data in Intrastat.

Table 2
Estimated trade without historical data 2005

	Non-response trade (SEK billion)	Below threshold trade (SEK billion)
Arrivals	8.5 (38 %)	14.0 (62 %)
Dispatches	4.2 (26 %)	11.9 (74 %)

From Table 2, differences can be seen between arrivals and dispatches for estimated data without history; 38 percent of arrivals relate to PSI's that rarely or never report their Intrastat figures, corresponding to 26 percent for dispatches. There are therefore more "problematic" non-response companies for arrivals than for dispatches.

If we relate our estimated data to the total value (collected + estimated), we get the following table:

Table 3
Total Intra trade divided into collected and specified estimated trade 2005

	Collected value (SEK billion)	Estimated with historical data (SEK billion)	Estimated without historical data non-response (SEK billion)	Estimated without historical data under threshold (SEK billion
Arrivals	532.2 (94 %)	13.2 (2 %)	8.5 (1.5 %)	14.0 (2.5 %)
Dispatches	542.3 (96 %)	8.1 (1.5 %)	4.2 (0.5 %)	11.9 (2 %)

From Table 3, we can see that the estimated value share in 2005 was 6 percent for arrivals and 4 percent for dispatches. The corresponding figures for the first half of 2006 (and the earlier publishing) are otherwise slightly under 8 percent for arrivals and 5 percent for dispatches.

During 2005 non-response amounted to roughly 1 300 arrivals companies and 450 dispatches companies every month (see Table 4). During the first half of 2006, the corresponding figures were roughly 1 800 for arrivals and 550 for dispatches.

Table 4
Estimated monthly average Intra trade divided into PSI's and non-PSI's 2005

	Estimated with historical data (number of PSIs)	Estimated without historical data (number of PSIs)	Estimated without historical data (number of non-PSIs)
Arrivals	566	735	23 000
Dispatches	262	172	11 000

Companies without history are divided into size categories (according to their annual VAT values), three categories for arrivals and three for dispatches (<SEK 10 million, SEK 10-100 million, >SEK 100 million). The division into size category in terms of value for March 2006 is shown in Appendix 1. It can be noted that the size categories are today not equal according to size. After consideration the size categories were amended as follows: SEK 0-4 million, SEK 4-40 million and >SEK 40 million. The change in the criteria for historical data meant that at least one month should exist from the past twelve month period in order for the history to be used, instead of from the past six month period, as previously. Table 5 shows the effect of the change in criteria for the month of March 2006. Additional months should be studied before any certain conclusions can be drawn.

Table 5
Impact of the change in criteria on the possibility of using historical data

Estimated with historical data (number of PSIs). Criteria: 12 months.		Estimated with historical data (number of PSIs). Criteria: 6 months.		Difference
arrivals (PSIs):	957	arrival (PSIs):	880	+ 77 (PSIs)
dispatches (PSIs):	334	dispatches (PSIs):	307	+ 27 (PSIs)
arrivals (SEKm):	2 318	arrivals (SEKm):	2 279	+ 39 (SEKm)
dispatches (SEKm):	1 116	Dispatches (SEKm):	1 042	+ 74 (SEKm)

The effect of the change in the criterion for history resulted in roughly 100 data providers, who previously had been estimated without history, being estimated with history. The total value for both the flows amounted to slightly over SEK 100 million. The total share of values estimated with history of the total non-response values increased from 63 percent to 76 percent.

If we compare the size categories before and after the changes (see Appendix 2), we can see that a certain share of the trade is now being estimated with history instead of using the method for those without history. The number of companies in the smallest group has decreased. Because the trade structure differs markedly between arrivals and dispatches companies, different group divisions should possibly be applied for the different flows. However, for reasons of simplicity related to the design of the estimation process and the effect of unique divisions on accuracy, we have chosen to use the same size divisions for both flows.

When the non-response application was created in 1999, when the size categories were also determined, the conditions were different than those today. Non-response was considerably greater and the threshold values have been significantly raised since then. More of the very large companies who were obliged to provide information were non-response and were placed in the third size category. In addition, there were serious problems in the beginning to find out the addresses of the foreign data providers. Intense non-response work is carried out today focusing primarily on the largest and most significant companies, resulting in a very low level of non-response in the largest group.

How great is the difference when comparing the current criterion of a historical period of six months and the new historical period of twelve months, at the same time as the estimation groups for companies without history are amended? The difference has been studied at commodity code level. Each commodity code has first been classified according to its annual value (SEKm) and then classified according to the absolute difference in percentage between the old and the new criteria based on data for March 2006.

In table 6, the commodity codes with annual values of at least SEK 100 thousand are shown, detailed in four different value groups. In addition, we compare the absolute percentage difference between the current criteria and the new criteria for arrivals and dispatches. In the table we can see that 991 commodity codes in the arrivals with a yearly value of >10 SEKm show absolute differences of between 0-25 percent when comparing the new criteria to the old.

The most interesting observation is of course for the largest commodity codes in terms of value (>SEK 10 million), where the difference is at least 75 percent. It can be seen that 25 commodity codes in arrivals and 63 in dispatches show a difference of at least 75 percent, at the same time as these are much larger in terms of value. It is clear that when estimating without history, we spread out the value over a large number of commodity codes. This means that, with the new criteria, we estimate fewer commodity codes but these are larger in terms of value. Roughly 500 commodity codes that were previously spread out disappear with the new criteria. Around 50 of these relate to commodity codes worth >SEK 10 million. It should also be noted that around 70 percent of the codes in arrivals and 75 percent of the codes in dispatches did not show any differences at all.

Table 6
Difference between new and old criteria in terms of number of commodity codes divided by size and annual value

Arrivals Annual value (SEKm)	Number of commodities Difference in percentage				
	0-25	25-50	50-75	75-100	Total
0.1 - 1.0	1 289	260	67	315	1 931
1.0 - 5.0	1 564	155	32	129	1 880
5.0 - 10.0	605	21	0	23	649
>10.0	991	8	0	25	1 024
Total	4 449	444	99	492	5 484

Dispatches Annual value (SEKm)	Number of commodities Difference in percentage				
	0-25	25-50	50-75	75-100	Total
0.1 - 1.0	913	115	86	215	1 329
1.0 - 5.0	852	118	70	128	1 168
5.0 - 10.0	292	26	6	54	378
>10.0	668	13	7	63	751
Total	2 725	272	169	460	3 626

It is thus clear that some differences do occur when we convert to using the new criteria. The question is how great the difference is compared to the actual value and how great the difference is in the revised values.

We have compared March 2006 at the first publishing date (47 days after the end of the month of March) with March 2006 three months later, regarding the original and the new criteria. Again we have studied data at commodity code level. We get the following table:

Table 7
Revised commodity codes using the old criteria in terms of number of commodity codes divided by size of difference and annual value

Arrivals Annual value (SEKm)	Number of commodities Difference in percentage				
	0-25	25-50	50-75	75-100	Total
0.1 – 1.0	1 148	314	113	356	1 931
1.0 – 5.0	1 498	202	55	125	1 880
5.0 – 10.0	597	33	4	15	649
>10.0	979	21	3	21	1 024
Total	4 222	570	175	517	5 484

Dispatches Annual value (SEKm)	Number of commodities Difference in percentage				
	0-25	25-50	50-75	75-100	Total
0.1 - 1.0	920	138	96	175	1 329
1.0 - 5.0	868	127	86	87	1 168
5.0 - 10.0	318	25	7	28	378
>10.0	691	29	9	22	751
Total	2 797	319	198	312	3 626

Table 8
Revised commodity codes using the new criteria in terms of number of commodity codes divided by size of difference and annual value

Arrivals Annual value (SEKm)	Number of commodities Difference in percentage				
	0-25	25-50	50-75	75-100	Total
0.1 - 1.0	1 468	94	40	135	1 737
1.0 - 5.0	1 631	63	23	57	1 774
5.0 - 10.0	607	14	4	10	635
>10.0	972	9	3	13	997
Total	4 678	180	70	215	5 143

Dispatches Annual value (SEKm)	Number of commodities Difference in percentage				
	0-25	25-50	50-75	75-100	Total
0.1 - 1.0	1 108	28	14	33	1 183
1.0 - 5.0	971	22	13	20	1 026
5.0 - 10.0	324	5	1	5	335
>10.0	660	22	5	2	689
Total	3 063	77	33	60	3 233

It can be seen that the new criteria seem to give better estimations at commodity code level, as the commodity codes do not need to be revised as much. The share of commodity codes in arrivals that needed to be revised by at least 50 percent decreased from 12.6 percent using the current criteria to 5.5 percent. In dispatches, this share decreased even more, from

14.1 percent to 2.9 percent. One should be cautious with any conclusions based on one month only.

It is perhaps most important to study the high-value commodity codes (>SEK 10 million) which have been revised by at least 50 percent. The greatest difference is also seen here in dispatches; 31 commodity codes were revised by at least 50 percent using the current criteria, compared to 7 commodity codes with the new criteria.

On chapter level (CN2 level), the number of chapters needing to be revised by at least 25 percent decreased by three chapters (from six to three) in arrivals. In dispatches the number of chapters decreased by four (from 7 to 3).

Table 9 shows the chapters that, using the current criteria, were revised by at least 25 percent and compared these to the revision necessary if we were using the new criteria.

Table 9
Chapters that were revised by at least 25 percent using the old criteria, compared to the size of the revision necessary using the new criteria (March 2006)

Flow	Chapter	Revision 1 (SEKm)	Revision 2 (SEKm)	Abs. diff
arrivals	14	0.6	0.6	0
arrivals	43	-0.4	-0.4	0
arrivals	47	74.7	-2.9	71.8
arrivals	78	-0.9	0.0	0.9
arrivals	80	1.4	0.1	1.3
arrivals	97	-3.7	-3.7	0
dispatches	6	3.9	0.0	3.9
dispatches	7	6.1	0.1	6.0
dispatches	12	50.4	50.0	0.4
dispatches	49	-28.2	-5.8	22.4
dispatches	67	1.6	0.0	1.6
dispatches	93	3.4	0.0	3.4
dispatches	97	2.4	0.2	2.2

For arrivals, we can above all note large differences in Chapter 47 (Fibrous cellulosic material) and in dispatches Chapter 49 (Printed books, newspapers, etc.). Chapter 12 (Oil seeds and oleaginous fruits) is however the chapter with the largest revisions, something which cannot be seen here.

On a three-digit SITC (Standard International Trade Code) level, the number of codes needing to be revised by at least 25 percent is unchanged for arrivals and four codes disappear completely. In dispatches, the number of codes decreases by 8 (from 18 to 10) and 5 codes disappear completely. The SITC codes in dispatches in which the difference appears to be greatest in terms of less revisions are: 591, 597 and 774. The SITC3 codes which disappear completely in arrivals are 072, 211, 325 and 343, and in dispatches are 044, 266, 281, 289 and 972.

3.3 Seasonal impact on calculations of distribution keys

For non-response companies, the total arrivals and/or dispatches are estimated using one of the methods described in Section 2.2. Companies under the threshold value are estimated using either data from VAT information for the month in question, method (*iii*), or using method *(v)*, see Section 2.2. The estimated values for arrivals and/or dispatches are divided into commodity codes and countries using the distribution keys based on the companies' previous trade or according to similar companies, see Section 3.1.

In order to estimate arrivals and/or dispatches divided into goods and countries, one of the methods described in Section 2.2 is used to estimate total arrivals or dispatches and the total trade is then divided into commodity codes and countries using one of the distribution keys in Section 3.1. The estimates of total trade are calculated with or without a seasonal component. These seasonal components relate to the company's industry. For companies that have previously reported trade, companies with history, the estimated trade is divided into commodity codes and countries using the previously reported values. Three months from the past six month period are used for the divisions and any seasonal patterns for the different commodity codes are not taken into account when dividing the total estimates. For companies without history, reported values from responding companies are used to divide the estimated total trade into commodity codes and countries; any seasonal patterns for the different commodity codes are therefore taken into account as the estimations are based on responding companies in the month in question.

For distribution keys estimated with history, seasonal patterns for the different commodity codes should also be studied and taken into consideration in the estimation. Estimating seasonal components for all commodity codes is probably not possible because there are a large number of commodity codes and the number of observations per commodity code is not sufficiently large for all commodity codes. It should be possible to estimate seasonal components for commodity codes on an aggregated level and use them when dividing trade into commodity codes. This means that the distribution key based on company history should be changed to take into account seasonal patterns for the different commodity groups.

3.4 Division of traders without historical Intra data

The groups are to a great extent created manually roughly every third year (see Section 3.1) and a large amount of resources are needed for this process. Furthermore, a great deal of subjective judgements are made when classifying the groups. The current group classification should be automated so that groups are updated every month.

Statistics Sweden's Business Register (FDB) contains information on the company's industry classification according to the Swedish Standard Industrial Classification, (SNI). A company's primary economic activity should at least give some information about the trade it carries out even if the actual trade in goods can differ significantly from the industry classification. A company can be more or less specialised and the trade with other member states can represent a greater or lesser part of the total economic activity. However, even if there are differences, they should be similar so that the industry classification gives valuable information when estimating the trade of companies without history. The industry classification used by the estimation system is not currently updated regularly and the routines for continuous updating of industry codes will be implemented. This means that the industry codes will be as up-to-date as possible.

Table 10
Number of estimation groups per flow and size category

Size category	Flow	Number of groups
0-10 SEKm	arrivals	26
10-100 SEKm	arrivals	20
>100 SEKm	arrivals	13
0-10 SEKm	dispatches	26
10-100 SEKm	dispatches	22
>100 SEKm	dispatches	23

According to Table 10, we have a total of 59 estimation groups in arrivals and 71 estimation groups in dispatches. Every size category and flow includes a "miscellaneous" group. The share of NACE codes that are placed in the miscellaneous group is divided according to the following six group levels:

Table 11
Share of 3-digit NACE codes in miscellaneous group

Size category	Flow	Share of NACE codes in miscellaneous group
0-10 SEKm	arrivals	0.708
10-100 SEKm	arrivals	0.762
>100 SEKm	arrivals	0.856
0-10 SEKm	dispatches	0.632
10-100 SEKm	dispatches	0.744
>100 SEKm	dispatches	0.874

Small dispatches companies show the lowest share of NACE codes in the miscellaneous group (63.2 percent) while the largest companies in dispatches show the greatest share of NACE codes in this group (87.4 percent). In total, there are 223 possible 3-digit NACE codes.

The criterion used currently in the (manual) groups compilations is that the SITC3 groups sorted by value in each NACE3 group should constitute at least 80 percent to be selected. Logical rationality of the compilations is then checked and incorrectly placed groups are removed. If the group, after this clean up of "incorrect" groups, still constitutes at least 60 percent of the NACE3 code's original value, or there are at least 5 companies, a separate estimation group can be created. In other cases, these companies are placed in the miscellaneous groups.

In table 12 estimated trade for non-respondents is divided by estimated value by ordinary groups and estimated value by miscellaneous groups. In table 13 estimated trade for companies under the threshold is divided in the same way as in table 12. The data used in tables 12 and 13 is from the production month August 2006.

Table 12
Estimated value being placed in normal groups and in miscellaneous groups, estimated value of non-response

Size category	Flow	Estimated value by ordinary groups (SEKm)	Estimated value by miscellaneous group (SEKm)
0-10 SEKm	arrivals	236	68
10-100 SEKm	arrivals	55	44
>100 SEKm	arrivals	190	21
Total:	arrivals	481	133
0-10 SEKm	dispatches	114	21
10-100 SEKm	dispatches	95	25
>100 SEKm	dispatches	117	14
Total:	dispatches	326	60

Table 13
Estimated value divided into normal groups and miscellaneous group under threshold value

Size category	Flow	Estimated value by ordinary groups (SEKm)	Estimated value by miscellaneous groups (SEKm)
0-10 SEKm	arrivals	1 243	416
10-100 SEKm	arrivals	8	5
>100 SEKm	arrivals	0	0
Total:	arrivals	1 251	421
0-10 SEKm	dispatches	1 160	161
10-100 SEKm	dispatches	19	18
>100 SEKm	dispatches	0	19
Total:	dispatches	1 179	198

The tables show that a large amount of the value that are estimated by non-response without history and for companies under the threshold value are found in the size category SEK 0-10 million.

We are therefore often talking about small data-providing companies or companies that are not obligated to provide data. The share of the trade that is estimated without history that refers to size categories SEK 0-10 million amounts to a full 84 percent. The share of the value of the non-response without history that is placed in the miscellaneous group is 19 percent. The share of the value of trade under the threshold value that is placed in the miscellaneous group is slightly over 20 percent. Almost 3.5 percent of the trade that is estimated without history relates to companies with an annual VAT value of >SEK 10 million; their trade is placed in the miscellaneous group. It is clear that the larger companies that are obligated to provide data are less likely to be non-respondents than the smaller companies, and historical data exist for these. It seems that the greatest problem is related to arrivals companies that are not obligated to provide data, where almost a half a billion SEK are placed in the miscellaneous group.

How can we automate the classification into groups?

Firstly, automatic routines for steps 1-3 should be implemented (see section 3.1) with the new size categories and criteria for historical data. Thereafter, the division shares for commodity groups in each industry classified NACE3 code could be used in an cluster analysis to create homogenous groups in every size category and flow.

Up to step 3 (see section 3.1), we have arranged data in the following way for each 3-digit industry classification code. Examples of withdrawals from NACE group 012 for arrivals are given below:

NACE3 = 012 (animal husbandry), arrivals

NACE3	VALUE	SHARE
001	14 793 862	0.42
034	8 028 539	0.20
721	6 424 266	0.17
291	5 166 247	0.16
:	:	:

For every size category, we can have a maximum of 223 groups, which is the total number of 3-digit NACE codes. Too many groups, making estimation in groups with a small number of companies will definitely be less accurate. Using the manual routines, the number of groups varied for every size category from between 13 to 26 groups, where the highest value size category had the lowest number of companies. The automatic classification method should give a fairly similar result.

Each NACE3-group has been given a distribution key with shares per three-digit NACE-code according to the actual trade. Each key is based on the yearly collected trade for the six size classes, three in arrivals and three in dispatches.

The distribution key is defined by

$$p_{ij} = \frac{y_{ij}}{\sum_j y_{ij}},$$

where y_{ij} is the trade value for the NACE3 group *i* for trade group *j*, and the trade group is defined as commodity code and country.

When the groups are to be created, we can use collected Intradata for 2005 at enterprise level. One problem that has occurred is that sometimes a corporate registration number exists without an industry code. This relates in total to 444 arrivals enterprises to a value of SEK 11188 m and 184 dispatches enterprises to a value of SEK 14755 m. In an attempt to reduce this classification non-response slightly, reclassification based on actual trade has been done. In order for reclassification to be done, the largest value NACE code at 3-digit level must constitute at least 60 percent. The criterion for every group created is that at least five enterprises must be included in the group. It can be interesting to study the division of the maximum 223 classified NACE groups for which there are not at least five enterprises. Tables 14 and 15 below show the data before and after reclassification. Arr1 =arrivals for the group of smallest companies and Dis3=dispatches for the group of largest companies and so on.

Table 14
Before reclassification of missing NACE-codes

	Arr1	Arr2	Arr3	Dis1	Dis2	Dis3
Number NACE3 <5 enterprises	117	120	119	119	123	116
Reclassified enterprises, number	291	110	52	87	64	41
Reclassified enterprises, value (SEKm)	776	1826	9229	271	1209	14980

Table 15
After reclassification of missing NACE-codes

	Arr1	Arr2	Arr3	Dis1	Dis2	Dis3
Number NACE3 <5 enterprises	114	116	115	117	120	116
Reclassified enterprises, number	139	57	34	30	25	16
Reclassified enterprises, value (SEKm)	413	956	6042	103	518	4124

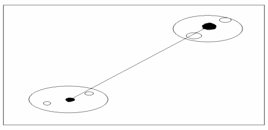
It can be seen from the tables above that reclassification does have an effect. A marked effect can above all be seen in the size category Dis3 with the largest dispatches enterprises, in which a trade value of close to SEK 15 billion before reclassification related to enterprises without NACE3 code, was reduced to SEK4 billion after classification.

Cluster analysis:
Cluster analysis is a method for grouping individuals, objects or variables into unknown groups. The cluster method is highly empirical. Different methods can lead to different groupings, both in number and content, and there are different methods to use within the cluster analysis. One commonly-used method in cluster analysis is the 'centroid method', using the mean or centroid values within each cluster when calculating the distance between the clusters.

The distance between two clusters, x and y, is defined as the (squared) Euclidean distance as $d(x , y) = | x - y |^2$.

In the centroid method, the distance between two clusters is defined as the (squared) Euclidean distance between their centroids or means.

The picture below illustrates the distance between two clusters when using the centroid method.



Initially according to earlier experience we decide to have 30 groups in size category 1, 25 in size category 2 and 20 in size category 3, i.e. 150 groups in total. This is done by the FASTCLUS procedure in SAS. Some of the text below is taken from the SAS-manual:

The FASTCLUS procedure combines an effective method for finding initial clusters with a standard iterative algorithm to minimise the sum of squared distances from the cluster means. The method is based on nearest centroid sorting. A set of points called cluster seeds is selected as a first estimate of the means of the clusters. Each observation is assigned to the nearest seed to form temporary clusters. The seeds are then replaced by the means of the temporary clusters, and the process is repeated until no further changes occur in the clusters.

The FASTCLUS procedure in SAS operates in four steps:

1) Observations called cluster seeds are selected.
2) When specifying the DRIFT option, temporary clusters are formed by assigning each observation to the cluster with the nearest seed. Each time an observation is assigned, the cluster seed is updated with the current mean of the cluster. This method is sometimes called incremental, online or adaptive training.
3) If the maximum number of iterations is greater than zero, clusters are formed by assigning each observation to the nearest seed. After all observations are assigned, the cluster seeds are replaced by either the cluster means or other location estimates (cluster centres) appropriate to the LEAST=p option. This step can be repeated until the changes in the cluster seeds become small or zero.
4) Final clusters are formed by assigning each observation to the nearest seed.

The initial cluster seeds must be observations with no missing values. The maximum number of seeds (and, hence, clusters) can be specified using the MAXCLUSTERS= option. A minimum distance by which the seeds must be separated can also be specified using the RADIUS= option.

The procedure always selects the first complete (no missing values) observation as the first seed. The next complete observation that is separated from the first seed by at least the distance specified in the RADIUS= option becomes the second seed. Later observations are selected as new seeds if they are separated from all previous seeds by at least the radius, as long as the maximum number of seeds is not exceeded.

If an observation is complete but fails to qualify as a new seed, PROC FASTCLUS considers using it to replace one of the old seeds. Two tests are made to see if the observation can qualify as a new seed.

Firstly, an old seed is replaced if the distance between the observation and the closest seed is greater than the minimum distance between seeds. The seed that is replaced is selected from the two seeds that are closest to each other. The seed that is replaced is the one with the shortest distance to the closest of the remaining seeds when the other seed is replaced by the current observation. If the observation fails the first test for seed replacement, a second test is made. The observation replaces the nearest seed if the smallest distance from the observation to all seeds other than the nearest one is greater than the shortest distance from the nearest seed to all other seeds. If the observation fails this test, PROC FASTCLUS goes on to the next observation.

Table 16 shows some details of the run of the cluster analysis.

Table 16

Details from the run of cluster analysis:

	Arr1	Arr2	Arr3	Dis1	Dis2	Dis3
Decided number of clusters	30	25	20	30	25	20
Number of iterations required	5	6	9	5	7	5
Criteria for initial seeds	1.242	1.392	1.238	1.341	1.294	1.333
Criteria for last seeds	0.031	0.032	0.028	0.037	0.039	0.044
Share of 3-digit NACE codes in the largest group	48 %	61 %	60 %	48 %	53 %	64 %

The R-square (explanatory degree) of each variable indicates how important the variable is for the cluster. The expected value for the overall R-square is under the uniform zero hypothesis, assuming that the variables are uncorrelated. The value is missing if the number of clusters is greater than one-fifth of the number of observations. To test the hypothesis that the cluster means are equal, the procedure runs a pseudo F statistics.

To test the separation among all clusters and test the numbers of chosen clusters, the pseudo F statistics can be used, where:

F-pseudo = (R-square/(C-1))/(1-R-square)),

c = the number of clusters

n = the number of observations.

Large F-values indicate that the number of chosen clusters is acceptable and that the separation among all clusters is high.

In addition, a procedure "proc candisc" was run. In canonical correlation analysis, we examine the linear relationships between a set of X-variables and a set of Y-variables. The technique consists of finding several linear combinations of the X-variables and the same number of linear combinations of the Y-variables in such way that these linear combinations best express the correlation between the two sets. Those linear combinations are called the canonical variables, and the correlations between corresponding pairs of canonical variables are called canonical correlations. It is necessary to analyse the canonical structure and also to test the zero hypothesis that the canonical correlation in the current row and all that follow are zero.

Multivariate test statistics, such as Pillai's Trace, Wilks' Lambda, Hotelling-Lawley Trace, and Roy's Greatest Root, are used in the analysis. The first three statistics are defined in terms of all the squared canonical correlations. Here, there is only one linear combination (the transformation) and therefore only one squared canonical correlation of interest, which is equal to the R-square. These statistics are normally defined in terms of the squared canonical correlations, which are the eigenvalues of the matrix H (H+E)-1, where H is the hypothesis sum-of-squares matrix and E is the error sum-of-squares matrix. Here the R-square is used for the first eigenvalue, and all other eigenvalues are set to 0 since only one linear combination is used. The tests are used for testing the hypothesis that the means of the classes of the selected variables are equal in the population.

Table 17

Analysis from group divisions by cluster analysis

Measurement	Arr1	Arr2	Arr3	Dis1	Dis2	Dis3
F-pseudo	10.31	10.31	13.09	9.40	7.03	8.04
Expected overall R-square	0.289	0.256	0.225	0.284	0.258	0.214
Number of variables (NACE3-groups) where R-square >0.6	91	75	56	90	41	44

In Table 17, we can see that the F-values and the explanation degree (R-square) are as a rule lower for dispatches than for arrivals. In the categories with the medium-sized and larger companies, in particular, these values appear to be lower. We can also see that there are considerably fewer NACE3 groups showing an explanation degree of at least 0.6 in these categories.

All categories show significance at the level p<0.0001 in the tests in Wilks' Lambda, Pillai's Trace, Hotelling-Lawley Trace and Roy's Greatest Root.

The twelve NACE3 groups 241, 275, 300, 322, 323, 341, 351, 372, 501, 611, 725, 922 showed an explanation degree of at least 0.6 in at least five of the six categories. NACE3 group 725 showed an explanation degree of at least 99 percent in all categories.

To compare the old manually-divided groups with the new groups that have been divided using the cluster analysis method, we can study the effect on the number of commodity codes for which at least 50 percent of the trade was estimated for the last published period, November 2006, with their five historical months.

The result is the following:

Old groups:
219 commodity codes for which at least 50 percent of the trade value for each commodity code was estimated.

New groups:
161 commodity codes for which at least 50 percent of the trade value for each commodity code was estimated.

It seems therefore that the new groups contribute to a lower number of commodity codes for which at least 50 percent of the value of the code was estimated.

An advantage is that the automatic method takes the whole distribution key for each 3-digit industry code into consideration while the manual method looks primarily at the largest commodity groups that are included.

In appendix 3 the old and new groups of the size class with the smallest companies in arrivals are compared. In the new groups the miscellaneous groups have been divided in two groups (number 9 and 99). In total 110 of the new groups were betrayed as miscellaneous. Of the old groups 158 groups were classified as miscellaneous.

3.5 New distribution level for total Intra trade

The estimated arrivals and dispatches divided into commodity codes and countries are only saved by flow, commodity code and country. Imputations on company level are not saved; when calculating total trade for a subgroup of all companies, e.g. companies of a certain size, it has so far been necessary to use the total collected values.

There is a need to be able to calculate totals, i.e. collected plus estimated values, for subgroups of companies.

When calculating total trade, collected values, imputed values for non-response and an estimation of the values under the threshold value are used. Imputations and estimations of the trade under the threshold value are made in the non-response application. Every time a month is published, new data are taken from the database. Non-response, measured in the number of companies, decreases over time so that, at the second publishing time for a month, the non-response is less than it was at the first publishing. New VAT data are available for every publishing time. New imputations for non-response companies and companies under the threshold value are calculated for every publishing time. An imputed value for a company at the first publishing time does not necessarily match the imputed value at the next publishing time. This is, among other things, because the number of companies included in the VAT data is larger at the second publishing time than at the first. Imputation for commodity codes and countries also differs at the different publishing times, as the estimation of distribution keys for companies without history is based on a different number of responding companies at different times.

In order to produce estimates of total trade for, for example, one commodity group for a subgroup of companies, it is necessary as previously mentioned to have collected or imputed values for the included companies. New imputations can be calculated every time a total estimation is to be produced or the imputations from the latest publishing time can be saved at company level and used for estimations of this kind. The former solution is preferable if late-arriving responses are to be included and to better utilise VAT data.

Small reporting groups can result in a large part of the value being estimated and the uncertainty in the estimations can be great. It is necessary to measure in some way which results are reliable. This applies naturally to all published figures but, if a subgroup of all companies is of interest, there is a significant risk that precision in the estimations will not be sufficiently good.

3.6 New routines for controlls of output data

With every monthly production, a supplementary SAS program is run for the follow-up of output data on the following levels:

a) Total follow-up

b) Follow-up commodity and country

c) Follow-up enterprise

In an ordinary month, there is a maximum of 16 percent unweighted non-response (number of companies) for arrivals and 13 percent unweighted non-response in dispatches. The weighted non-response (sum of value) rarely exceeds 6 percent for arrivals and dispatches. The coverage non-response in value should not be greater than 3 percent for arrivals and dispatches. An investigation should be carried out if any of these conditions are not met. It may be necessary to carry out more comprehensive correction work followed by a new production run.

Total follow-up (a) relates to the follow-up of total non-response and trade under the threshold value per flow for the revised time periods.

Follow-up commodity and country (b) relates to different commodity group codes within SITC, NACE and CN for which more than 50 percent of the value of the commodity codes and 20 percent of the value of the country codes has been estimated and ranked by the share of estimated trade.

Follow-up enterprises (c) relates, among other things, to the follow-up of non-response enterprises that are obligated to provide data and for which there is no VAT data, but that have been estimated with a VAT value (estimated or actual value) or non-response enterprises for which more than 50 percent has been estimated.

Extra output controls for the distribution keys are not currently carried out but should be implemented in the long-term.

There are some ideas to implement controls on and continuous follow-up of larger revisions. We could, for example, look at the ratios between estimated and reported value per country and chapter in total or divided into the different cases:

1) Non-response divided with history
2) Non-response divided without history
3) Trade under the threshold value with history
4) Trade under the threshold value without history

The total of the absolute differences (noted as Δ) between the estimated value and the reported value on different aggregation levels can be expressed as:

$$\Delta = \sum_{j=1}^{J} \left| \sum_{i} \left[\left(\hat{p}_{ijm} - p_{ijm} \right) \sum_{j=1}^{J} y_{ijm} \right] \right|$$

where p_{ijm} is the distribution key for enterprise *i*, aggregation level *j*, month *m*, $\hat{p}_{ijm}$ is the estimated distribution key and y_{ijm} is the company's reported trade with *j*, month *m*. For every commodity group (or country), the difference between the estimated trade and the actual trade in month *m* is calculated and Δ is obtained by summing the absolute sums of these differences.

6 Future developments at SCB

In connection with the ideas that have been presented in this project, primarily with the aim of improving the accuracy of the estimated Intrastat trade, the following proposals for project activities have been put forward for 2007 and 2008.

1) Changes in the criteria from 6-month historical data to 12-month data in the non-response application.
2) Implementation of routines to classify enterprises that lack an industry code based on actual trade.
3) Changes of size categories and implementation of new estimation groups where there is no history in the non-response application.
4) Implementation of supplementary output controls of revisions with regards to the distribution keys used to divide the estimated trade.
5) Evaluation of the estimation methods for total estimated trade per company.
6) Implementation of additional methods to those used when there is no history to obtain information from the enterprises that disappear from the Intrastat system when the threshold value is raised.
7) Changes in the non-response application to make it possible to produce total Intrastat data on enterprise level divided by commodity code and country.
8) Implementation of new tables in MSSQL in which there should be updated total enterprise data divided by commodity code and country.
9) Improvement of follow-up reports and clarification of the content in connection with the production runs.
10) The industry classification used by the estimation system is not currently updated regularly and the routines for continuous updating of industry codes will be implemented.
11) Evaluation of the effects of the proposed improvement measures for the full 2006 Intrastat data.

Appendices

Appendix 1A

Old size classes for the estimated trade, where historical trade is missing- non-PSI data for March 2006

FLOW	SIZE CLASS	MONTHLY VALUE (MKR)	SHARE (%)
arrivals	0-10 mkr	1443	99%
arrivals	10-100 mkr	7	1%
arrivals	>100 mkr	0	0%
	Sum:	**1450**	**100%**
dispatches	0-10 mkr	1275	98%
dispatches	10-100 mkr	23	2%
dispatches	>100 mkr	0	0%
	Sum:	**1 298**	**100%**

Appendix 1B

Old size classes for the estimated trade, where historical trade is missing- PSI data for March 2006

FLOW	SIZE CLASS	MONTHLY VALUE (MKR)	SHARE (%)
arrivals	0-10 mkr	386	57%
arrivals	10-100 mkr	112	16%
arrivals	>100 mkr	180	27%
	Sum:	**678**	**100%**
dispatches	0-10 mkr	110	21%
dispatches	10-100 mkr	288	56%
dispatches	>100 mkr	119	23%
	Sum:	**517**	**100%**

Appendix 2A

New size classes for the estimated trade, where historical trade is missing- non-PSI data for March 2006

FLOW	SIZE CLASS	MONTHLY VALUE (MKR)	SHARE (%)
arrivals	0-4 mkr	1392	96%
arrivals	4-40 mkr	43	3%
arrivals	>40 mkr	15	1%
	Summa:	**1450**	**100%**
dispatches	0-4 mkr	1182	91%
dispatches	4-40 mkr	104	8%
dispatches	>40 mkr	13	1%
	Summa:	**1299**	**100%**

Appendix 2B

New size classes for the estimated trade, where historical trade is missing- PSI data for March 2006

FLOW	SIZE CLASS	MONTHLY VALUE (MKR)	SHARE (%)
arrivals	0-4 mkr	332	49%
arrivals	4-40 mkr	258	38%
arrivals	>40 mkr	88	13%
	Summa:	**678**	**100%**
dispatches	0-4 mkr	62	12%
dispatches	4-40 mkr	233	45%
dispatches	>40 mkr	222	43%
	Summa:	**517**	**100%**

Appendix 3

Comparison of old and new estimation groups – Arr1

Group number	Old groups	New groups
1	011,524,801	353,621
2	013,851	365,923,927
3	014	223,334,714,921
4	020	300,671,722,723,726,922 924
5	050	232,241,243,246,284
6	151,526,713,722,726,731	181,191
7	152,512,513,521,522,523	314,321,323,527,746
8	153,503,504,511	050,152,745
9	156	012,013,014,015,020,103 112,120,131,142,157,159 174,182,192,193,203,205 211,212,222,242,251,252 263,264,265,266,272,282 283,285,286,287,291,293 294,295,296,311,312,313 315,316,331,332,333,342 343,351,361,363,364,366 401,402,403,410,453,454 502,503,511,514,515,518 519,524,525,526,552,554 602,603,622,623,634,641 651,652,660,672,701,702 712,741,742,743,744,747 748,751,752,753,801,802 803,853,900,911,912,930 950,990
10	158,721,748,922	354,504
11	177,182	352,601
12	193	132,244,851,852
13	203,204,205,211,212,361 501,502,515	141,267
14	221,222,252	275,292,335,355,455,631 732
15	241,245	143,268,274,362
16	246,251	145
17	273	341,501,505,711,926
18	281,315,742,744,803	281,451,713,725
19	334	233,261,262,297,372
20	353,743	011,019,151,153,154,155 156,158,512,513,521,522 553,555
21	514	271,273,371,452
22	518,527,741	201
23	519	202,204
24	702	322,642,721,731
25	900,926	724

(forts.)

Group number	Old groups	New groups
26		160,175,247,551
27		611,612,632,633,703
28		171,172,176,177
29		000,245,523
30		173,221,804,913,925
99	012,015,019,101,102,103 111,112,120,131,132,141 142,143,144,145,154,155 157,159,160,171,172,173 174,175,176,181,183,191 192,201,202,223,231,232 233,242,243,244,247,261 262,263,264,265,266,267 268,271,272,274,275,282 283,284,285,286,287,291 292,293,294,295,296,297 300,311,312,313,314,316 321,322,323,331,332,333 335,341,342,343,351,352 354,355,362,363,364,365 366,371,372,401,402,403 410,451,452,453,454,455 505,525,551,552,553,554 555,601,602,603,611,612 621,622,623,631,632,633 634,641,642,651,652,660 671,672,701,703,711,712 714,723,724,725,732,745 746,747,751,752,753,802 804,852,853,911,912,913 921,923,924,925,927,930 950,990, 000, missing codes	101,102,111,144,183,231 missing codes

2004:5	Designutredning för KPI: Effektiv allokering av urvalet för prismätningarna i butiker och tjänsteställen. Examensarbete inom Matematisk statistik utfört på Statistiska centralbyrån i Stockholm
2004:6	Tidsserieanalys av svenska BNP-revideringar 1980–1999
2004:7	Labor Quality and Productivity: Does Talent Make Capital Dance?
2004:8	Slutrapport från projektet Uppsnabbning av den ekonomiska korttidsstatistiken
2004:9	Bilagor till slutrapporten från projektet Uppsnabbning av den ekonomiska korttidsstatistiken
2004:10	Förbättring av bortfallsprocessen i Intrastat
2004:11	PLÖS. Samordning av produktion, löner och sysselsättning
2004:12	Net lending in the Swedish economy. Analysis of differences in net lending National accounts (NA) – Financial accounts (FA). Background – comparisons - analysis
2004:13	Testing for Normality and ARCH. An Empirical Study of Swedish GDP Revisions 1980–1999
2004:14	Combining leading indicators and a flash estimate
2004:15	Comparing welfare of nations
2004:16	ES-avdelningens utvecklingsplan 2004
2004:17	Den svenska konsumentprisindexserien (KPI), 1955–2004. En empirisk studie av säsongsmönstret. En tillämpning av TRAMO/SEATS
2004:18	Skola, vård och omsorg i privat regi. En sammanställning av statistik
2005:1	An ignorance measure of macroeconomic variables
2005:2	Svenska hälsoräkenskaper. Ett system framtaget inom ramen för de svenska nationalräkenskaperna
2005:3	The sample project. An evaluation of pps sampling for the producer and import price index
2005:4	Finansiellt sparande i den svenska ekonomin. Utredning av skillnaderna i finansiellt sparande. Nationalräkenskaper, NR – Finansräkenskaper, FiR. Slutrapport
2005:5	Net lending in the Swedish economy. Analysis of differences in net lending National accounts, NA – Financial accounts, FA. Final report
2005:6	Varför får NR motstridiga uppgifter? Statistikens samanvändbarhet studerad inom ramen för Nationalräkenskaperna
2005:7	Evaluation of the possibility of producing statistics on production in the construction industry on a monthly basis
2005:8	Increased Automation of the Validation and Correction Processes in the Swedish Intrastat Production
2005:9	The Growth Rate Method for Production of Rapid Estimates of the Swedish Foreign Trade – Development, Evaluation and Preparations for the Implementation
2005:10	Svenskt producentprisindex (PPI) 1975–2004. En analys av tidsseriens integrationsgrad och säsongsmönster
2005:11	Kvaliteten på Sveriges statistik över import och export – En normativ och kvantitativ studie
2005:12	Implementing Simplified Intrastat Reporting in Statistics Sweden - Evaluation and Preparations for a Future Implementation
2005:13	FoU-verksamhet i kommuner och landsting – en pilotstudie inför kommande undersökningar
2006:1	Den svenska regionala tillväxten av i dag – en undersökning med utgångspunkt i "Den Nya Ekonomin"
2006:2	En flashestimator för den privata konsumtionen i Sverige med hjälpvariablerna HIP och detaljhandeln – En tidsserieanalys med hjälp av statistikprogrammet TRAMO
2006:3	A Selective Editing Method considering both Suspicion and Potential Impact, developed and applied to the Swedish Foreign Trade Statistics
2006:4	Rapport från projektet PLÖS II. Förslag till förbättrad samordning av kortperiodisk statistik avseende produktion, löner och sysselsättning
2006:5	Skola, vård och omsorg i privat regi. En sammanställning av statistik 1995–2005
2007:1	Intrastat system – Presumption for One flow system
2007:2	Revisions of production data in the Swedish national accounts

ISSN 1650-9447 (Print)
ISSN 1654-0719 (Online)

ISBN 978-91-618-1374-2

Statistikpublikationer kan beställas från SCB, Publikationstjänsten, 701 89 ÖREBRO, e-post: publ@scb.se, telefon: 019-17 68 00, fax: 019-17 64 44. De kan också köpas genom bokhandeln eller direkt hos SCB, Karlavägen 100 i Stockholm. Aktuell publicering redovisas på vår webbplats (www.scb.se). Ytterligare hjälp ges av Bibliotek och information, e-post: information@scb.se, telefon: 08-506 948 01, fax: 08-506 948 99.

Statistical publications can be ordered from Statistics Sweden, Publication Services, SE-701 89 ÖREBRO, Sweden (phone: +46 19 17 68 00, fax: +46 19 17 64 44, e-mail: publ@scb.se). If you do not find the data you need in the publications, please contact Statistics Sweden, Library and Information, Box 24300, SE-104 51 STOCKHOLM, Sweden (e-mail: information@scb.se, phone: +46 8 506 948 01, fax: +46 8 506 948 99).