Increased Automation of the Validation and Correction Processes in the Swedish Intrastat Production

2005:8

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

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

Increased Automation of the Validation and Correction Processes in the Swedish Intrastat Production

2005:8

Statistics Sweden
2005

Economic Statistics

Increased Automation of the Validation and Correction Processes in the Swedish Intrastat Production

Statistics Sweden
2005

Tidigare publicering Publicerad årligen sedan 2001
Previous publication Previous publication has been made since 2001

Producent SCB, Avdelningen för ekonomisk statistik
Producer Box 24 300
104 51 Stockholm

Förfrågningar Anders Jäder, tfn: +46 8 506 947 28
Inquiries anders.jader@scb.se

Omslag: Ateljén, SCB

ISSN 1650-9447

Printed in Sweden
SCB-tryck, Örebro 2005:02

Preface

This report concludes the project Increased Automation of the Validation and Correction Processes in the Swedish Intrastat Production, which was conducted during January-August 2004. The report presents the investigations and the analyses carried out within this project and the proposals made by the project group to decrease the manual work and to increase the level of imputation. The results of this project will facilitate the implementtation of increased automation in the validation and correction processes.

The project has been financed by the EU with the funding from the EDICOM 2004 budget.

This report has been produced by the participants of the project group, Anders Jäder, Ulrica Häll and Martin Fors, at the unit of Foreign Trade.

Statistics Sweden, April 2005

Kaisa Ben Daher
Head of Foreign Trade Unit

Abstract

Control and correction of Intrastat data is both time and resource consuming. In order to decrease the manual work associated with control and correction procedures this project aims to increase the level of imputation in Intrastat. The focus is on decreasing manual work related to errors with small impact on the published statistics in order to be able to concentrate human resources to errors with large impact.

For the validation process, a test is performed to see whether the current thresholds used for imputation could be increased to allow for more imputations. It is found that a moderate increase in the thresholds is possible without damaging the quality of the published statistics to a large extent. This approach would result in a decrease in the workload of 550 observations per month.

For the unit price checking it is tested whether it would be possible to decrease the number of edited observations by 700 lines a month from the current 1500 lines a month. The results indicate that the decrease would to some extent deteriorate the quality of the published statistics. Whether this deterioration is acceptable or not is open for debate.

We also investigate the interaction between the validation process and the unit price checking. We propose further work in order to harmonize the methods used in the two processes.

Statistics Sweden, April 2005.

Contents

Summary and conclusions

In this report an effort is made to find methods to decrease the manual work associated with the production of Intrastat statistics. The focus is on increasing the automation in the process for low value transactions in order to concentrate the manual work on transactions that have a large impact on the published data. Methods are proposed and evaluated for the validation and unit price checking processes since these processes require the highest amount of manual work.

The report is divided into three sections: The validation process, the unit price checking process and the interaction between the validation and unit price checking. The last section stems from the fact that imputed values are flagged in the unit price checking, thereby causing manual work. Moreover, changes in the number of imputed values will therefore have an effect on the unit price checking.

In the section for the validation process we study and test five different methods, one of them being a test that tries to replicate the method currently used. All of the methods are based on using different threshold values for observations to be deleted or imputed. One of the methods, which can be compared to a totally automated validation process is evaluated. This method has the thresholds for imputing values set at their maximum value and the deviation of the variable Total Invoiced Amount from the sum of the reported values is allowed to be 999 per cent. This method is considered to produce data of poor quality since it allows even high value transactions to be imputed, therefore no manual correction would be done for these observations.

The method that is considered to decrease the manual work the most without lowering the quality of the published data uses threshold values that are approximately twice the size of the threshold values today. Also the threshold for deleting observations is increased, from 6,000 SEK and 5 kilos to 20,000 SEK and 10 kilos. The allowed deviation for Total Invoiced Amount is doubled from 10 to 20 per cent. When using this method the amount of observations requiring manual validation would decrease by approximately 550 observations per month.

Also a number of other proposals are given, although not evaluated. These proposals concerns observations with missing values for all variables that should be deleted automatically, furthermore small negative values should be deleted by same method as positive values. Also a proposal is made to impute the country codes D and GE with DE since this is most often the intended country code. These proposals do not make a large effect on the number of manually corrected observations, but it still seems justified to make these changes in the system.

We have also evaluated some techniques to decrease the manual work involved in the unit price checking. Our first approach is to test what effect the unit price checking has on the published statistics and whether the effect is small enough to make the whole process unnecessary. When evaluating the effect on invoiced value it turned out that corrections of on

average 797 million SEK is made for arrivals and on average 395 million SEK for dispatches. Some of these corrections cancel each other out, so the effect on total arrivals and total dispatches is about 100 million lower. Evaluations on SITC 1-digit level showed differences between the unedited and edited value in per cent of the manually edited value of up to 27 % and differences in SEK as high as 2,500 million SEK. On CN6 level differences in per cent of the manually edited value can become several thousand per cent and the differences in SEK can become up to 266 million SEK. Even thought this is not our decision, it seems that these errors cannot be accepted. Some unit price checking has to be done.

We proceeded in the report by evaluating what the effect on the published statistics would be if the unit price checking was decreased by 700 observations every month, from 1,500 to 800 observations. For arrivals for the reference month June 2004, a graph (Figure 2) is presented that shows the effect on the total value as a function of the number of edited observations. It shows that the effect on the total value diminishes gradually with the number of edited observations.

All evaluation concerning the unit price checking has this far been made for the variable invoiced value. This is partly because it is probably the most important variable and partly because it is much more easy to evaluate the effect on invoiced value than on other variables. Aggregations of weight and supplementary unit are less informative than aggregations on value. However, graphs (Figure 3 and Figure 4) are presented in the report where the differences in weight and supplementary unit have been transformed into differences in value. The graphs show the effect of the unit price checking as a function of the number of edited observations. When comparing Figure 2 with Figure 3 and Figure 4 it is clear that much of the effect of the unit price checking is on weight and supplementary unit. This can also be seen from the fact that the hit-ratio is higher for both weight and supplementary unit than for invoiced value.

For the variable invoiced value evaluations were made on commodity code level. Compared to the scenario where no observations are edited the approach with 800 edited observations seems much better. However some errors are still corrected because of the editing of the last 700 observations. Only 3 of the SITC1-codes published for the first 6 months of 2004 would have had errors exceeding 1 % of the fully edited value. On CN6-level the remaining errors are more visible. Some CN6-codes have large differences expressed as per cent.

Whether a decrease is possible in the number of edited observations in the unit price checking is not up to us to decide. If the quality of invoiced value on CN6-level seems reasonable in the scenario where 800 observations are edited, it must still be remembered that the quality on weight and supplementary unit must be taken into account. The quality of these variables is more difficult to assess. It is interesting to note that after 1,500 observations have been edited the overall hit-ratio is still around 40 %.

The last part of the report is devoted to the interaction between the validation process and the unit price checking. It is found that, on average, 113 imputed observations are flagged in the unit price checking every month. If the imputation is increased according to the moderate approach described above the number of flagged observations will increase by 7

observations per month. The increase is not large but it is certainly a step in the wrong direction. On the other hand, if the use of the moderate approach is combined with a decrease in the number of edited observations in the unit price checking the number of flagged imputed observations will decrease to 108 per month, which after all, is still unsatisfying.

To cope with the problem, the rules for imputation in the validation process should be harmonized with the unit price checking. We have tested a very simple version of this. The prices used for imputation, which are calculated by country, are replaced with prices from the unit price checking calculated regardless of country. The effect is small on the validation process but the flagging of imputed observations is decreased by 12 observations per month. Further harmonization, e.g. imputation by PSI where possible, would most certainly give further decreases in the number of flagged observations.

In the report proposals are also given for further studies as well as proposals for changes in the IT-system, which would facilitate evaluations in the future.

1 Introduction

1.1 Background

Control and correction of Intrastat data is a time and resource consuming activity. Some of the validation procedures at Statistics Sweden are already automated, but still much control and correction work is done manually. In the validation process, which is one of the two most resource consuming processes, the automated procedures are applied for the items not having large values using threshold levels. The variables corrected automatically are commodity codes from 4-digit level to 7-digit level, country codes and transaction codes. Both missing codes and invalid codes can be handled in the automated procedure. Automatic correction also applies for missing information on net mass, supplementary units and invoiced value for low value items. The other major resource consuming process is the unit price checking. This process is at present not automated at all.

1.2 Implementation timetable for the operation

The operation was started in January 2004 and completed in August 2004.

1.3 Objectives

The objective of this action is to investigate possibilities to increase the level of automation regarding transactions with relatively low values (increase the threshold values for automated corrections). The aim is to be able to concentrate manual control and correction work on significant errors (high value transactions). The aim is also to decrease the human resources currently used for control and correction work.

1.4 Human resources used

The total time required to carry out the operation is estimated to be 700 hours. Approximately 100 hours will be used for administration and translation. The remaining 600 hours will be distributed equally on the three participants of the work group:

Martin Fors (200 hours)
Ulrica Häll (200 hours)
Anders Jäder (200 hours)

1.5 Equipment and software applications or programmes used

For evaluation studies the software SAS has been used. The unit price checking application is also constructed in SAS. A test database is used, which is a copy of the Intrastat production system. In this test environment, different tests are made.

1.6 Description of the operation

The original objective of the operation was to find ways to increase the level of automation in the Intrastat production process. In the report it has been found that in the validation process it is possible to increase the level of automation by increasing the threshold values, without lowering the quality of the data. This would decrease the number of observations that require manual work with approximately 550 observations every month. The evaluation of the validation process is given in chapter 3.

Furthermore, in chapter 4, the unit price checking has been studied in order to evaluate the possibilities to decrease the number of checked observations from 1,500 to 800. The results indicate that using this method might deteriorate the quality of the published statistics.

Finally, in chapter 5, studies are made regarding the interaction between the validation process and the unit price checking. This is due to the fact that it has been discovered that observations that are imputed during the validation process still have to be checked during the unit price checking. A proposal is made to harmonize the rules for imputation in the validation process with the unit price checking.

The above results were found by comparing the results from using the different methods with the results generated by the currently used method. During the operation some problems were discovered regarding the possibilities to evaluate the effect of the new methods in the validation process. The problems concern the current IT system, which uses different primary keys in different tables making it difficult to trace a specific observation through the production process. In order to facilitate evaluations in the future a proposal has been made to introduce a transaction number for each observation, that follows the observation through the production process.

In the report proposals for further studies are given. It is proposed that the possibilities to impute missing commodity and country codes should be studied. One way of doing this is to impute all variable values of the observation with values from a similar observation. Further harmonization between the validation process and the unit price checking is also recommended. For the unit price checking it is proposed that the possibilities to impute values should be investigated. This requires knowledge on which variable that is erroneous (value, weight or supplementary unit). Preliminary studies indicate that this might be possible in some cases.

2 Intrastat at Statistics Sweden

In this chapter an overview of the Intrastat system at Statistics Sweden is given. This is followed by a section where estimates on the human resources used in the production process are presented.

2.1 General description

The purpose of the Foreign Trade statistics at Statistics Sweden (SCB) is to illustrate Sweden's trade in goods with the rest of the world. The published statistics includes two different surveys, Intrastat (trade with EU countries) and Extrastat (trade with countries outside the EU, "third countries").

Two publications are released each month. A preliminary publication is released about four weeks after the reference month. This release only contains total exports, total imports and net value of trade. About ten weeks after the reference month, the detailed statistics are published. In this publication the distribution of Foreign Trade on countries and commodities is presented. Data from the detailed statistics are also used for the publication of volume indices, Foreign Trade in constant prices, which are published on a quarterly basis.

Data for Extrastat is collected from the Swedish Customs. The data is based on the customs declarations from the importing or exporting company. Also, before the material is delivered, a control and correction procedure takes place at the Swedish Customs. Since only output checking is done at SCB, the following text will focus on Intrastat.

The Intrastat survey has been conducted since Sweden joined the EU in 1995. Until 1994 data was collected from the customs declarations. The Swedish Customs carried out the survey until 1999, but since 2000 Statistics Sweden is responsible for the survey.

The sample in Intrastat consists of around 15,000 companies (Providers of Statistical Information, PSIs) that are selected using a cut off procedure, the respondents together report around 350,000 items to Intrastat every month. All companies that have annual arrivals and/or dispatches above 1.5 million SEK are required to leave reports to Intrastat. If they fail to do so, they can be ordered to pay a fine in accordance with Swedish laws. Information about total arrivals and dispatches for the companies is gathered from their tax declaration via cooperation with the Swedish tax authorities.

The information from the tax declarations is also used in the control and correction work as well as adjusting for non-response and lack of coverage. During 2003 the lack of coverage was for arrivals around 1.8 per cent, and for dispatches around 0.8 per cent of the value of the total Foreign Trade. At the same time the non-response rate was two per cent of the value of dispatches and three per cent of the value of arrivals.

In the middle of every month questionnaires are being sent to PSIs that use paper questionnaires for their reports. This is the largest media with around 140,000 observations reported every month. The PSIs also have the possibility to leave an electronic report, via the systems EDI or IDEP. Around 100,000 observations are reported on each of these two medias every

month. Since May 2004 the PSIs also have the possibility to use a web questionnaire, although only around 500 observations was reported this way during the first month that the questionnaire was in use. Nil reports are possible to leave via TDE, Touchtone Data Entry.

The electronic reports are registered and validated at the time of delivery, which means that the respondent is not able to deliver a report with invalid data. This is not the case for paper questionnaires, which usually arrive at SCB around two weeks after the end of the reference month. The questionnaires are then sent to a company that register the information on all questionnaires, this procedure takes about one to two weeks. When the information is registered and returned to SCB the process of validating the information begin.

The validation takes place during two weeks, starting about four weeks after the end of the reference month. During the validation the reports are automatically controlled to ensure that they contain all necessary information. If an observation misses commodity code, country code, weight, supplementary unit or invoice value the system marks the observation as invalid. This is also the case if a report contains values that cannot be correct, for example a non-existing country or commodity code.

When the validation is completed the unit price checking and total PSI value checking starts, about six weeks after the reference month. During the price control an automatic system studies kilo prices and unit prices. Prices that have a large deviation from historic prices and/or have a large impact on the published figures are signalled as possible errors. In the company control the total values from the reports are evaluated. This is done using historic information as well as information from the tax decalrations.

Almost two months after the reference month all the control and correction work is finished and the publication process starts. A more detailed description of the validation process is given in chapter 3 and for the price checking process in chapter 4.

2.2 Human resources used in the production process

The production work in Intrastat includes distribution and collection of questionnaires (both paper and electronic questionnaires), registration, helpdesk, checking procedures, correction of data and, to some extent, development work.

A total of 18 persons are working with the production of Intrastat, of these 6 have a university degree and 12 are clerks. Out of the 18 persons working with Intrastat production, 11 work full-time with Intrastat production, the remaining 7 persons devote part of their time to other tasks such as development work, dissemination etc.

According to previous calculations the staff working full-time with the production devote around 60 per cent of their working time to checking and correction of data. This means that approximately six or seven full-time workers are required to manage the checking and correction of data. Besides the checking on micro level two or three employees work approximately three days each every month with checking on macro level.

For comparison estimated work devoted to some of the other tasks is as follows. Work related to helpdesk (assisting PSIs) requires two full-time workers, distribution and collection of paper questionnaires requires approximately three full-time workers. Work related to electronic data collection and registration requires one full-time worker each.

The above-mentioned figures do not include human resources relating to IT or methodology issues.

3 The validation process

In this chapter the validation process is studied in order to try to increase the level of automation. The first part of the chapter describes the current validation process. Following this is a study of the manual and automated corrections made during the validation process in June 2004. Some variables are chosen to be evaluated and are then studied more carefully. Methods for increasing the number of imputations are described in part 3.3, these methods are finally evaluated in the last part of the chapter.

3.1 Description of the validation process

The validity checks made by the system when a report is inserted into it are described in this section.

The reported variables are checked in the following order:

- Reference period
- VAT-number and Subsidiary reporting number
- Report number
- Arrivals/Dispatches
- Total number of items
- Sum of invoiced value
- Item number
- Commodity code
- Country code
- Weight
- Supplementary unit
- Invoiced value
- Transaction Code
- Mode of transport

The methods for imputation of commodity code, country code, weight, supplementary unit and invoiced value will be described in this section. Comments are also made on the checks of total number of items and the total invoiced value.

There are several ways to replace erroneous commodity codes with an imputation. If the commodity code is not valid for the reference year but it is valid for the previous year the commodity code is transformed into a commodity code that is valid for the reference year. How the transformation should be done is specified in a control table, which has to be updated every year. At the turn of the year one code can be divided into two or more new codes and also two or more codes can be merged into a single code. Other possibilities exist as well. Eurostat send out information about these changes every year. How the replacement of codes should be done is not obvious, some manual work is involved.

If the given commodity code is valid for neither the reference year, nor the previous year, the system tries to impute a commodity code by using the first 7 digits of the given code. If for example the stated code is 8456 30 12, a code

which is not valid for the reference year 2004, it can be replaced by one of the valid codes 8456 30 11 or 8456 30 19. If there exist no codes on the stated 7-digits, the system tries to impute by using the first 6 digits in the same manner. If this is not successful the system tries to impute from the first 4-digits. There are often many possible codes to choose from when imputing. Which code to choose must be specified in a control table. The aim is that the largest code measured in value should be the code chosen, this table also has to be updated every year. For each type of imputation (4-, 6- or 7-digit level) there exist threshold values over which no imputation is done.

The variable country code is also checked. It should of course be non-missing and referring to a country that was a member of the European Union during the reference month. Some common errors are often made by the PSIs. Often for example the country code UK is given instead of GB. These errors can be corrected automatically by using a control table. No threshold is used for the correction. Instead all codes are corrected where possible. The control table has to be updated every year as well although it can simply be replicated from previous year if no changes are necessary. The table used for 2004 can be seen below in Table 1.

Table 1
The table currently used to convert specific erroneous country codes into new correct country codes

Erroneous country code	New country code
AU	AT
BL	BE
EL	GR
IR	IE
NE	NL
SF	FI
SP	ES
TY	DE
UK	GB

Sometimes the PSIs send in country codes referring to a country outside the European Union. It was thought previously that we could simply delete these lines automatically (with the exception of some codes which are similar to country codes in the European Union). The computer system has therefore been prepared to allow for this. However, the opinion among the clerks working with the correction of codes was that quite often it is found out that the observation should be reported after all, even though for a different country. That is why this possibility has not been exploited.

For the imputations of net weight, supplementary unit and value a price register is used. The prices in the price register are computed from previously reported observations for 12 month. The 12- month period is from two months before the last published month and twelve months back. Prices per kilo and prices per supplementary unit are calculated by commodity code and, if possible, by commodity code and country code.

The next variable after country code to be checked is net weight. The net weight must exist if it is compulsory for the commodity code. If it is not compulsory it is nevertheless not deleted. If net weight is missing for a compulsory code the system tries to impute it. The imputation is only done if

there exist a stated invoiced value on which the imputation can be based. The stated invoiced value must also be below the threshold value set for imputation. The imputation is done, if possible, by using the price per commodity code and country. Otherwise it is done by using the price per commodity code if such a price exists. The imputed weight is computed as the stated invoiced value divided by the price per kilo.

The supplementary unit must exist for commodity codes where supplementary unit is compulsory. If it is stated on a commodity code for which it is not compulsory, it is deleted. In cases where the commodity code is in error any stated supplementary unit is kept and if no supplementary unit is stated the system does not try to impute any. In the same way as for net weight, no imputation is done if an invoiced value is not provided. To be able to impute a supplementary unit the stated value must also be below the threshold value. The imputation is performed using the price per commodity code and country if possible, otherwise the system tries to impute using the price per commodity code. The imputed supplementary unit is the stated invoiced value divided by the price per supplementary unit. If the value zero is imputed, it is changed to 1.

The invoiced value must be non-missing. If it is missing the system tries to impute it by using the stated net weight and the price per kilo for the commodity code and country or just for the commodity code. If the net weight is not compulsory or it has not been stated the system tries to impute using the supplementary unit if it is compulsory and has been stated. After the imputation of the invoiced value has been done it is checked whether the imputed value is below the threshold set for the imputation.

Finally, if there are still errors, it is checked whether the observation can be erased. This is only done if the invoiced value is below the stated threshold value and the net weight is below the stated threshold for net weight.

For paper reports, the PSIs are obliged to fill in the total invoiced value of the whole report in the beginning of each report. It is checked whether the stated total invoiced value is equal to the actual sum of the reported observations. If the deviation between the two is larger than 10% of the stated total value the report is considered to be in error and it is further checked.

For paper reports, the PSIs are also obliged to fill in the total number of items on their report. It is then checked whether this number is equal to the actual number of items on the report.

3.2 Manual corrections in the current validation process

In order to evaluate for which variables it is possible and meaningful to increase the number of imputations, we have studied the total number of errors and the number of errors corrected manually for each variable. The study was made on data regarding June 2004. Some of the variables are not included here since there where no errors for these variables during this month[1]. The result is illustrated in Table 2 below.

[1] These variables are Report Number, Arrivals/dispatches, Form, Period, Agent's VAT number and Agents subsidiary number. During 2004 these variables has had none or very few errors.

Table 2
Description of errors and manual corrections in Intrastat for June 2004

Variable	Number of errors	Manual corrections	Manual corrections in %
Transaction Type	23,566	0	0.0
Item Number	1,940	1,939	99.9
Enterprise	373	371	99.5
Subsidiary Number	4,644	4,199	90.4
Total Invoiced Amount	22,235	11,020	49.6
Total Number of Items	16,132	11,567	71.7
Country Code	854	534	62.5
Commodity Code	6,134	1,182	19.3
Net Mass	1,796	268	14.9
Supplementary Unit	2,582	88	3.4
Invoiced Value	642	265	41.3
Total	80,898	31,433	38,9

As can be seen in the table, for the variable Transaction Type no manual work is required since all of the errors are corrected automatically. All errors are imputed with Transaction type=1, purchase/sale of goods, since this is the most common transaction type.

For the variable Item Number almost all errors are corrected manually. The main part of these errors is due to the fact that the variable's maximum value is 999. Thus, for some reason no more than three digits are accepted by the computer system. If a report has more than 999 items the enumeration starts over again from item number 1. In June 2004 there where only 28 enterprises that had an error for the variable Item Number. Furthermore, the manual work for correcting these errors is not very time-consuming since you can correct many errors with a small amount of work. Also no contact with the enterprise is required to make these corrections. Although the amount of work associated with these corrections is quite small, it would be desirable if we could avoid these problems since it is not an actual error. Efforts should be made to resolve the problem, either by changing the registration procedures or allowing for more digits in the item number.

The variable Enterprise also has a large proportion of manual corrections, although the total number of errors is quite small. The main part of the errors depends on the fact that there is no Enterprise stated on the report (286 out of 373 errors). It is almost impossible to automate the correction of the errors in this variable. One alternative could be to give all the reports with errors in Enterprise a simulated number. The downside of this is that we lose the possibility to control whether or not the enterprise is obliged to leave reports to Intrastat. It therefore seems as if it is difficult to increase the level of imputation.

For Subsidiary Number there were a total of 4,644 errors. Of these errors 445 where eliminated since the observation was considered to be less important, i.e. the value and the net weight was low. The remaining part of the errors was corrected manually. It is almost impossible to increase the level of imputation for this variable since it is an identification variable.

There were 395 reports that had a stated total invoiced value that was either missing or deviated more than 10 per cent from the sum of the invoiced values of the reported items. This corresponds to 22,235 individual observations. Approximately 50 per cent of these items were corrected manually. But, as in the case with Item Number, correction of one report leads to correction of a large number of items. Most of these errors are due to miscalculations or that the variable has been left out. Another reason for errors in this variable is typing errors during registration.

For the variable Total Number of Items the number of errors were in June 16,132, out of which 11,567 where corrected manually. As in the case above, correction of one report leads to correction of a large number of items. In this case the number of enterprises corrected manually were 136.

In June 854 errors were found for the Country Codes. More than 60 per cent of these errors had to be corrected manually. The main part of the errors is due to missing values, but it is also common that a country outside the EU is given or that only one letter is stated. For observations where the variable contains some information it might be possible to impute a correct country code. Currently some erroneous country codes are imputed with correct codes, for example the country code UK is automatically changed to GB. It is possible that the same solution can be used for other erroneous codes.

For the variable Commodity Code there where 6,134 items that contained errors. Approximately 20 per cent of the errors had to be corrected manually. The method for correcting errors in Commodity Codes is already quite good since it uses a lot of information but perhaps it could be improved to decrease the number of manual corrections.

For the variable Net Weight, there where 1,796 missing values in June 2004. The main part of these values was imputed using information from other variables and 268 of them had to be corrected manually. Although the proportion of automatically corrected errors is quite large there might be possibilities to increase the automatic corrections.

The number of imputations for Supplementary Unit is also high, of the 2,582 missing values only 88 had to be corrected manually. The main part of these observations seems to be otherwise correct so perhaps it would be possible to increase the level of imputation further.

The variable Invoiced Value has 642 errors during this month, but almost 265 of these have to be corrected manually. This corresponds to a bit over 40 per cent of the total errors and perhaps it could be justified to try to increase the level of imputations.

Based on the above information it was decided that the project should focus on increasing the level of imputation only for some of the variables. The chosen variables are: Country Code, Commodity Code, Net Weight, Supplementary Unit and Invoiced Value. In the following section a more thorough study is made for these variables. The study is made only on those observations that were erroneous and required manual correction since they could not be automatically imputed. The study is extended to include all of the first six months of 2004 to ensure that June was not an exceptional month.

In Table 3 below the number of manually corrected errors are illustrated for the first six months of 2004. Since it is common that the number of errors is larger during January a separate column with errors for January 2004 is included. The reasons for the large number of errors in January are mainly changes in the nomenclature regarding commodity codes, supplementary units etc.

Table 3
Manually corrected errors for the first six months of 2004

Variable	Manual corrections 200401–200406	Of which 200401
Commodity Code	7,201	1,182
Country Code	4,727	534
Weight	2,031	268
Supplementary Unit	1,104	88
Invoiced Value	1,733	265
Total number of errors	16,796	2,337
Total number of lines with minimum one error	14,506	3,057

The observations that could not be automatically imputed almost always have missing values in weight, supplementary unit and invoiced value. In some cases there has been non-numeric values given, but this is rare. In the variables commodity code and country code there are many different errors. In the process the system divides the errors according to Table 4 below. The number of errors for January 2004 is also shown to detect possible errors because of the change of year.

Table 4
Manually corrected errors in the Country and Commodity code for the first six months of 2004

Variable	Type of error	Number of errors	Of which 200401
Commodity Code	Missing	2,235	418
	Non-valid	4,544	675
	Code valid previous year	422	416
Country Code	Missing	2,569	467
	Non-valid	492	70
	Non-EU country	1,666	150

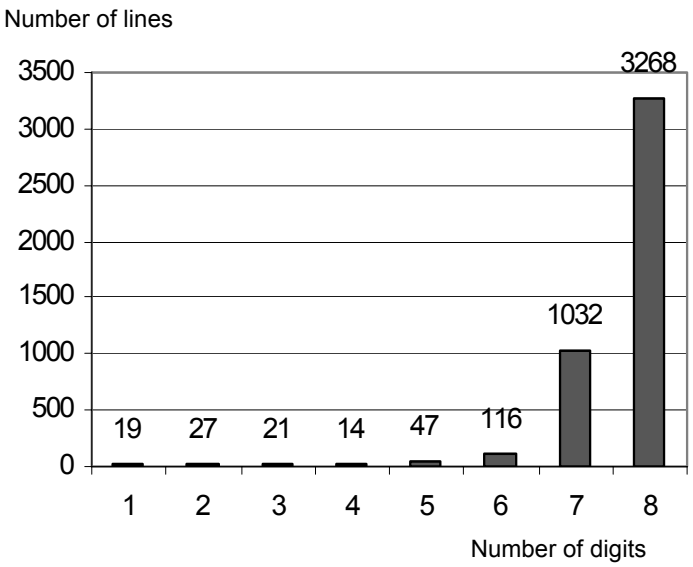
In the table, "missing" means that the code has not been stated by the PSI. Non-valid commodity code means that the commodity code is valid for neither the actual year nor the previous year, while "Code valid previous year" means that the commodity code was valid the previous year but is not valid the actual year. "Non-valid" for country code means that the code is not valid for any country, EU or non-EU. "Non-EU country" means that the code is valid, but not for a EU country.

Almost every error "Code valid previous year" that was not automatically corrected occurred in January. One of the reasons that such a large number of errors were not imputed was probably a delayed update of the control table that is used for the imputation. Almost half of the errors in the error type "Non-EU country" occurred for the reference month April 2004, most

probably because of confusion among the PSIs regarding the expansion of the EU that took place in May 2004.

Non-valid commodity codes can vary in the number of digits stated by the PSIs. This is illustrated in figure 1 below. It seems like most of the commodity codes do contain the correct number of digits, eight, but they do not make a valid code for some reason.

Figure 1
Number of digits on manually corrected non-valid commodity codes. Data for January to June 2004



After this overview of all of the errors it is now justified to study the errors more closely. This is done in order to find out why the errors are made and why it was not possible to correct them automatically. Because of the large number of data and more important, because of limitations in our data system when it comes to the possibilities of tracing observations, only a sample of the errors that were not imputed has been studied.

3.2.1 A sample-based study of errors that lead to manual corrections

A sample was drawn from data for the first six months of 2004 for each variable and type of error described above. The samples were studied and the results are described below.

For missing country code it seems as if the largest problem with this variable is that some enterprises do not state the country code at all. In some cases the code seems to be systematically left out. These errors represent a large part of the missing country codes, it therefore seems as if a lot could be won by contacting these enterprises on an early stage and try to make them leave correct reports. The reason that these errors are not automatically corrected is that no routine for imputation exists in the current validation process. Furthermore a large part of the observations containing missing country codes have invoiced values and weights above the threshold for automatic erasure.

When the country code is invalid the most common error seems to be that the PSIs state a country code consisting of only one letter. The most common

case is when the letter D is stated, the main part of these was changed to DE even though some were changed to DK. A routine for automatic correction of non-valid country codes exist, but only for specified codes. The above-mentioned example can be included in this group of codes.

Country codes belonging to a non-EU country is also a common error. It was studied if there are non-EU country codes that are similar to EU country codes thereby causing these errors. Some codes are similar to more than one EU country code, for example DM, LI and SL. For some codes it could be possible to impute a valid code instead of the erroneous code. This can be the case for GE, for which 18 out of 20 errors in the sample were changed to DE.

It is common that countries in Europe, but not a member of the EU, are stated. For example it is common that CH, Switzerland, and NO, Norway, are stated as country codes. It is also quite common that SE, the country code for Sweden is stated. Most of these errors in the sample were deleted in the manual correction.

When analysing the non-valid commodity codes the most common error seems to be that the external firm registers a correct commodity code incorrectly. Some errors are also due to incorrect codes stated by the PSIs. The reason for not imputing or deleting non-valid commodity codes is in almost every case that the first four digits do not create a valid CN-4 code and that the weight and invoiced value are above the limits for erasure. Some of the incorrect CN-8 codes do make up a correct CN-4 code, but the invoiced value is above the threshold value.

Concerning the cases when the commodity code is missing there is no function to impute a correct commodity code at the moment. The only method available is to erase the whole observation. The studied errors are all above the threshold values for erasure.

The codes valid the previous year could be imputed more often if control tables in the validation system were updated in time. Making sure that the non-valid commodity codes are replaced with new ones could save a lot of manual work especially in the beginning of every year.

In the variable net weight the most common errors are that the variable is left out or that a negative weight is stated. These errors are often not imputed because the invoiced value is also left out or negative. It is also common that the value is above the threshold value for imputation.

The majority of manual corrections in the variable supplementary unit are due to the variable becoming obligatory for the commodity code since the previous year. At first this makes imputation impossible since no previous observations exist on which unit prices can be based. The need for manual corrections decreases during the year, when reported data can be used to

calculate unit prices to be used in the price register. To illustrate this the distribution of the errors in the sample over months is shown in Table 5 below.

Table 5
Number of manually corrected errors in supplementary unit by reference month

Period	Frequency
200401	19
200402	17
200403	4
200404	5
200405	2
200406	3

For the variable invoiced value the most common errors are negative values are stated or that no value is stated at all. The reason they have not been imputed is that the weight and/or supplementary unit are also negative or missing. Many of the negative values are erased during the manual correction.

3.3 Methods

In order to find methods to decrease the level of manual work in the validation process we have focused on increasing the thresholds for imputation as well as the thresholds for erasure of observations. In order to do this, the Intrastat data for the first six months of 2004 are copied to test database. Thereafter it is possible to change the thresholds and replicate the validation process.

Five different methods for increasing the level of automation are tested. The methods are numbered 1 to 5, and a short description of the different methods is given below. For a complete description of the threshold values, see Annex 1.

Method 1, Current method
This method is the method that is currently used in the validation process. The threshold for deleting observations is 6,000 SEK and 5 kilos. For most types of errors the threshold values are below 1,000,000 SEK. The only errors that have thresholds above this value is when Supplementary Unit is missing or the Commodity Code is not valid this year (but was the previous year). In these two cases the threshold is as high as 10,000,000 SEK. For the variable Total Invoiced Value the deviation from the sum of the reported items is allowed to be 10 per cent.

The reason we test this method is that we want to see if there are any effects caused only by the change from the actual production environment to the test environment. The result of this method will be used as a basis with which the other methods are compared.

Method 2
This method has been used to be able to study the effect of a totally automated validation system, at least for imputations. Using this method, all threshold values for imputation are at their maximum levels. The deviation

for Total Invoiced Value is still not allowed to be more than 10 per cent. Observations are deleted using the same values as before (6,000 SEK and 5 kilos).

Method 3
The only difference from Method 2 is that the deviation of Total Invoiced Value from the sum of the reported items is allowed to be 999 per cent.

Method 4
Method 4 has been chosen to be able to evaluate the effects of erasing a larger number of observations. In this method the thresholds for imputation are at the same level as in the current validation process but erroneous observations that has a value below 20,000 SEK and weight below 10 kilos are deleted.

Method 5
Method 5 is an attempt to find a more useful combination of thresholds. In this method almost all threshold values for imputation are doubled. Some of the values are very high already in the current validation process and are therefore left at their original level. Also the allowed deviation for the Total Invoiced Value from the sum of the reported items is doubled (to 20 per cent). Furthermore, the threshold value for deleting items is kept at the increased level used in and Method 4.

Other proposals for increasing automation
During the project it has been discovered that, besides the proposed methods above, there are some minor changes that can possibly decrease the manual work associated with the validation process.

In the current computer system observations with missing values for all variables are included in the validation. These observations are most likely deleted by the clerk that validates the data and thus could be deleted before the manual validation.

Furthermore it is quite common that the PSIs report negative values instead of correcting earlier reports, causing the data to be validated. Unlike reported positive values there is no limit for when an erroneous observation should be deleted. It is reasonable to use the same limit for negative values as for positive values (6,000 SEK and 5 kilos). Another reason for this is that the Intrastat handbook states that corrections below 5,000 SEK in some cases do not need to be reported to Intrastat.

Another proposal is for the variable Country Code. When the errors made for this variable was studied it was discovered that it is sometimes quite obvious what country the PSI intended. Unfortunately some of the Intrastat Country Codes are quite similar and therefore it can be hard to find a solution on how to impute the values. For two erroneous codes we still find it justified imputing a correct value, at least for small values. This is when the PSIs report country code “D” or “GE”. When reports with this error were studied it was discovered that almost all errors were corrected to “DE” for Germany.

The proposals mentioned in this section are not evaluated since the effect on the final estimates is believed to be very small.

3.4 Evaluation of the methods

Before beginning the actual evaluations of the methods a short discussion is given below concerning what defines a good system of automatic imputation and how different systems can be compared with each other in order to find the best system.

The first requirement of an automated correction system is of course that it should produce variable values that are valid (e.g. that all commodity codes exists in the list of valid commodity codes and that all value variable values contains numeric values). It is easy to check whether this requirement is fulfilled and the answer is simply yes or no.

You could also reason that the automated correction system should create observations that are similar to the observations created by the manual correction system. One then considers the manual correction to generate the "true values" and the automated correction system should strive at coming as close as possible to these values. The closer a method of automated correction comes to the manually corrected values the better it is.

Another way of seeing it is that the manual process does not create true values. The manual process creates values stated by the PSIs but these values will then be checked for unreasonable values in the three checking processes Unit price checking, Total PSI value checking and Output checking. It is only after these checking stages that the true values are found. The automated values should thus, according to this view, try to come as close as possible to the values finally published.

Which view is the best is not obvious. It is possible that the result of the credibility checks is changes of a kind that these could impossibly be foreseen in a systematic way by an automated correction system. Which method comes closest to the final values becomes more a matter of chance than anything else. The effect of the credibility checks may also be so large that the difference between a manual and an automated correction system becomes insignificant in comparison.

Another indication of a good system of automatic correction is that it is in harmony with the unit price checking. If possible the criteria used to decide what is acceptable in the unit price checking should also be used to impute missing or incorrect variable values. In that way the automatically imputed observations will not require action in the unit price checking.

During the work of this project it has been discovered that there are some problems related to evaluating the validation process. This stems from the fact that different primary keys are used in different tables. Tracing observations through the system is therefore not always possible. In our report we have to some extent been able to avoid this problem by using our test database. By entering the same observations as in the ordinary production process into our test database we can e.g. evaluate different levels of imputation thresholds. However, to make this kinds of tests take some time and therefore another solution, which would enhance the possibilities for continuous quality control, would be preferred in the long run.

Although there might be some issues regarding the evaluation an effort has been made to try to evaluate the methods described in section 3.3. Since the validation process only applies for observations reported on paper questionnaires, only these observations have been studied. The evaluation has been

made for those observations that has been changed or erased due to the method used. The number of imputed and deleted observations has been studied, since the main objective is to decrease the amount of manual work. Further, the amount of increased automation for the variables Net Weight and Invoiced Value are studied. In 3.4.1 the effect of increasing the number of deleted and imputed observations is studied. A more detailed evaluation is done in the following section, 3.4.2, where also the effects on less aggregated levels (SITC 1-digit level and CN 6-digit level) are studied.

3.4.1 Overview of the magnitude of imputations and deletions

One way of decreasing manual work in the validation process is to increase the limit for deleting observations. Table 6 illustrates the number of observations deleted for each method. In the first row the total number of deleted observations is stated. The following rows give the number of deleted errors by variable. Note that the sum of all errors by variable not necessaryly is the same as the total number of deleted observations. This is due to the fact that an observation can have errors in more than one variable.

Table 6
Number of deleted observations by method. Period: 200401-200406

Variable	Method 1	Method 2	Method 3	Method 4	Method 5
Total	1,462	1,485	1,544	2,266	2,289
Country Code	587	592	603	888	893
Commodity Code	913	930	975	1,437	1,453
Net Weight	20	21	21	25	26
Supplementary Unit	212	212	215	308	308
Invoiced Value	28	30	33	35	37

As can be seen in the table methods 2 and 3 do not increase the total number of deleted errors compared to the current method (method 1), at least not to a large extent. Method 4 on the other hand increases the number of deleted items by almost 55 per cent. In this method the limit for deleting observations is raised to 20,000 SEK and 10 kilos. In the current validation observations below 6,000 SEK and 5 kilos is deleted. Using Method 5, which increases the thresholds for imputation as well as for deleting observations, increases the number of deleted observations just a little bit more. By using this method instead of the current method around 138 observations less, in average, have to be corrected manually every month.

In Table 7 below the total value for the deleted items is given by variable. It is clear that the total value excluded from Intrastat increases rapidly with the number of deleted items. Using Method 4 increases the value by approximately 240 per cent compared to the current method. Method 5 increases the value a bit more; the excluded value now 245 per cent higher. The largest increase can be seen for the variable Commodity Code where the value for the deleted observations increases by a little more than 250 per cent.

Table 7
Values in SEK millions for deleted observations by method.
Period: 200401-200406

Variable	Method 1	Method 2	Method 3	Method 4	Method 5
Total	2.10	2.13	2.22	7.20	7.27
Country Code	0.80	0.80	0.83	2.72	2.73
Commodity Code	1.36	1.38	1.44	4.72	4.77
Net Weight	0.01	0.01	0.01	0.04	0.04
Supplementary Unit	0.28	0.28	0.28	0.83	0.83
Invoiced Value	0.02	0.02	0.02	0.04	0.06

Also the weight in tonnes is affected by increasing the number of deleted items. In Table 8 the effect on this variable is illustrated. The effect on weight is not as large as the effect on the value. When comparing the current method with Method 4 it can be seen that the total weight that is excluded from Intrastat increases by around 180 per cent. The result is approximately the same for method 5.

Table 8
Weight in tonnes for deleted observations by method. Period: 200401-200406

Variable	Method 1	Method 2	Method 3	Method 4	Method 5
Total	2.11	2.13	2.20	5.88	5.92
Country Code	0.88	0.88	0.90	2.39	2.40
Commodity Code	1.28	1.30	1.35	3.66	3.68
Net Weight	0.02	0.03	0.03	0.05	0.06
Supplementary Unit	0.31	0.31	0.31	0.73	0.73
Invoiced Value	0.03	0.03	0.03	0.06	0.06

Another way of decreasing the manual work associated with the validation is to increase the number of observations that are imputed. In Table 9 below, the number of imputed observations is stated for each of the methods. Using the current method, Method 1, a total of 47,969 observations were imputed. Since the sum of imputations for each variable is more than 10,000 higher, it is obvious that around 20 per cent of the observations have errors in more than one variable.

When increasing the threshold values to their maximum, as in Method 2, the number of imputed observations is 50,848, which is 6 per cent, or 2,900 observations, higher than before. This can perhaps be considered to be a quite small increase regarding the fact that the threshold values are at their maximum level. Furthermore, for the variables studied, there is no significant increase in the number of imputed variable values. By also increasing the threshold level for Total Invoiced Value, allowing it to deviate with 999 per cent from the sum of the reported values (Method 3), the number of imputed observations is just below 53,000.

When using a method combining more modest increases in the threshold values with a higher value for erasing observations, as in Method 5, the result is approximately the same as when using the maximum values in Method 2. This could be an indication that some manual work can be avoided by increasing certain threshold values and maintain a good quality in the data.

Table 9
Number of imputed observations by method. Period: 200401-200406

Variable	Method 1	Method 2	Method 3	Method 4	Method 5
Total	47,969	50,848	52,941	48,540	50,471
Country Code	1,618	1,660	1,714	1,627	1,667
Commodity Code	24,839	26,384	27,335	25,171	26,157
Net Weight	8,892	9,874	10,254	9,091	9,766
Supplementary Unit	22,402	23,329	24,063	22,753	23,412
Invoiced Value	895	1,086	1,398	899	1,014

In Table 10 the changes in imputed values are illustrated. With the current method, Method 1, the total value of observations with at least one imputed variable is around 3.9 billion SEK. Increasing the thresholds to their maximum level increases the total value to almost 6.8 billion SEK. By also increasing the allowed deviation in Total Invoiced Value the total value increases to 8.4 billion SEK. Method 4 leaves the imputed value almost unchanged, which is not surprising since the only change in the method compared to the current method is an increased threshold for deleting observations. A more interesting finding is that Method 5, although increasing the number of imputed variables, do not increase the imputed value by more than 11 per cent.

A comment regarding the values for Method 3 needs to be made. This method allows a large deviation of the Total Invoiced Value from the sum of all reported items. This causes the imputed value for all the variables to increase, especially the variable Invoiced Value. The reason for this is that for reports with this type of error all reported items needs to be validated.

Table 10
Values in millions SEK for imputed observations by method. Period: 200401-200406

Variable	Method 1	Method 2	Method 3	Method 4	Method 5
Total	3,943.37	6,799.97	8,377.44	3,955.97	4,380.24
Country Code	588.53	604.58	610.23	595.20	597.18
Commodity Code	2,080.04	3,913.13	4,065.34	2,082.61	2,404.51
Net Weight	191.83	733.29	742.55	194.90	273.55
Supplementary Unit	1,658.24	2,540.80	2,776.94	1,660.19	1,770.09
Invoiced Value	15.25	245.29	1,572.39	15.30	28.16

As can be seen in Table 11 also the weight is affected by increasing the number of imputations. For the method with maximum values the weight for imputed observations increases from 390,385 tonnes to 745,077 tonnes. When allowing a large deviation for total invoiced value the weight increases a bit more, to 773,461 tonnes. As expected, method 4 gives an almost identical outcome as the current method. Method 5 increases the imputed weight to around 424,677 tonnes.

Table 11
Weight in tonnes for imputed observations by method. Period: 200401-200406

Variable	Method 1	Method 2	Method 3	Method 4	Method 5
Total	390,384.82	745,076.60	773,460.59	390,513.35	424,677.45
Country Code	20,862.92	21,498.17	21,547.35	20,886.78	20,964.59
Commodity Code	300,476.79	539,851.87	547,858.41	300,491.39	330,252.25
Net Weight	7,700.14	98,516.50	98,834.53	7,751.64	13,110.90
Supplementary Unit	190,175.95	269,582.26	271,730.45	190,207.99	199,262.86
Invoiced Value	909.47	18,914.59	38,392.36	945.23	1,526.81

To summarize the discussion above, it seems as if using Methods 2 or 3 would probably generate data with poor quality since a large part of the data will not be manually validated at all. For observations with large values it is most certainly justified to continue with manual validation. Using Methods 4 and 5 seem to generate more modest changes in imputations and deleted observations. Therefore it seems reasonable to continue the evaluation of these two methods further.

3.4.2 The effect on published statistics

The following evaluations have been made on manually validated data that has not been checked in the unit price checking process. This is done in order to evaluate the methods without the effects that the price checking process generates. Due to various reasons explained in it is difficult to trace observations through the system since different primary keys are used in different tables. However, by using our test database we can compare tables that have the same primary key. There is no need to trace observations from the input process to the throughput process. Instead we compare different throughput tables with each other; one table from the regular Intrastat production process and five other tables from the Intrastat test database. This procedure is not free from trouble but it avoids some of the problems.

An effort has been made to match observations from the tests with manually validated data from the Intrastat system. When doing this approximately 3,000 observations from the tests cannot be found in the Intrastat system. There can be various reasons for this. There is of course the possibility that these observations exist in the system, but that the matching was unable to be performed. Furthermore it is possible that the PSIs have sent a new report with corrections. Another possibility is that a clerk has deleted the observations during the process. When studying some of the observations that are not matched with data from the Intrastat system it seems as if all of the above is represented in the material. The major part of the observations seem to have received new report numbers, making it impossible to match the data.

In order to have control over the evaluated observations and knowing that we do only include data that belong in the analysis only observations that can be matched from the two tables are included in this section. Furthermore, as explained earlier, only Method 4 and 5 and the test of the current validation method (Method 1) are evaluated.

Evaluation on aggregated data

When evaluating the methods we have studied the values generated by each method only for observations treated by the method. These values are then

compared to the values for the same observations generated by the actual production process measured before the unit price checking starts.

Ideally the difference between the test of the current method (method 1) and the values generated by the actual production process measured before unit price checking should be zero. As can be seen in Table 12 and Table 13, where the differences are illustrated for arrivals and dispatches respectively, this is not true. The reason for this is that there might occur manual corrections that are not related to either one of the processes. Another reason is that the price register used in the actual production process is constantly changing whereas this is not the case in our test.

Table 12
Difference between the methods tested and the actual production process. Arrivals. SEK

	Current method	Method 4	Method 5
January	251,773	251,773	-64,415
February	-98,090	-117,379	373,402
March	-661,522	-690,941	-18,375
April	-22,512	-22,512	220,420
May	-8,003	-8,003	-229,118
June	-123,608	-122,212	-607,212
Average	-110,327	-118,212	-54,216

Table 13
Difference between the methods tested and the actual production process. Dispatches. SEK

	Current method	Method 4	Method 5
January	-759,507	-759,507	-313,552
February	-198,513	-198,513	-379,739
March	-7,782	-7,782	-410,938
April	-496,474	-496,474	-520,416
May	1,839	1,839	-562,588
June	5,689	5,689	500,718
Average	-242,458	-242,458	-281,086

For a fair evaluation though, it is of more relevance to see how the new methods (4 and 5) perform compared to the test of the current method rather than compared to the actual production process. In Table 14 and Table 15 the effects of introducing either one of these two methods are illustrated.

In Table 14 the results for arrivals are illustrated. As expected Method 4 does not change the imputed values to any large extent, since this method only changes the thresholds for deleting observations. More interesting is to study Method 5 where most of the thresholds are doubled leading to approximately 2,500 fewer observations to validate manually. The method changes the total value for arrivals with, on average 56,111 SEK. For individual months the differences are larger. For example for March 2004 the

value of arrivals would increase by 643,147 SEK, which can be considered to be quite small.

Table 14
Difference between Methods 4 and 5 and the test of the current method. Arrivals. SEK

	Method 4	Method 5
January	0	-316,188
February	-19,289	471,492
March	-29,419	643,147
April	0	242,932
May	0	-221,115
June	1,396	-483,604
Average	-7,885	56,111

The results for dispatches are illustrated in Table 15 below. The results are approximately the same as for arrivals. Method 4 does not change the imputed value at all. Method 5 on the other hand generates some deviations. The average deviation during the first six months of 2004 is –38,628 SEK. Also for dispatches there are large differences between the months. For May 2004 the total value of dispatches would decrease by 564,427 using this method and for June 2004 the value would increase by 495,029 SEK.

Table 15
Difference between Methods 4 and 5 and the test of the current method. Dispatches. SEK

	Method 4	Method 5
January	0	445,955
February	0	-181,225
March	0	-403,156
April	0	-23,942
May	0	-564,427
June	0	495,029
Average	0	-38,628

When studying the effects on the total value for arrivals or dispatches both method 4 and 5 seem to perform quite well. The differences from the current method are acceptable, although noticeable, for method 5. For method 4 there are almost no deviations from the current method when looking at the imputed value. However, to evaluate the performance of a method it is important to also study how well it works on a less aggregated level. In the following section a study is made of how well method 5 works for disaggregated levels, SITC1-level.

Evaluation on SITC 1-digit level
To evaluate the performance of method 5 on less aggregated levels the following study has been made. Method 4 is not considered in this section since this method does not change the values to any large extent. The data that has been validated according to method 5 are included and the differrences between this method and manually corrected data are calculated. The observations included are then grouped according to SITC on the 1-

digit level. In order to evaluate the impact of the differences they are then compared with the edited value for each SITC-group. Note that the edited value in this case is both validated and checked during unit price checking.

In Table 16 the differences for arrivals are illustrated on each SITC group. As can be seen in the table, no large differences in per cent have been found on this level. Looking at differences in value, the largest positive difference occurs at SITC 7, the difference being 1.92 million SEK between method 5 and manually edited data. The largest negative difference in value occurs for SITC 6, where the difference is -1.89 million SEK. These differences might seem large, but when comparing them to the total value for the group they represent 0.01 and -0.03 per cent respectively. Instead, when looking at the largest differences in per cent of total value, the following is discovered. The largest positive difference in per cent is 0.07 and the largest negative difference in per cent is -0.05. These results occurred for SITC 2 and SITC 8.

In Table 17 the differences for dispatches are illustrated in the same way. It can be noted that the differences for some of the groups are larger in per cent of total value than for arrivals. For SITC 4, which has the largest negative difference, the difference is -0.73 per cent of the total value for the group. This is a larger deviation than could be seen for arrivals, but it must still be considered to be quite small. The largest positive difference in per cent occurs for SITC 1 where the difference is 0.07 per cent of the total value for the group. The largest positive difference in SEK is for SITC 5 where the difference is 2.26 million SEK, which represents 0.04 per cent of the value for this group. The largest negative value in SEK also represents 0.04 per cent of the value for the group, SITC 8, and is approximately –1.6 million SEK.

From the above discussion it seems clear that if you are only interested in good quality of the value on a quite aggregated level, as SITC 1-digit level, method 5 seems to produce data with sufficient quality. Although, it is still of importance to investigate the method further to see how it affects the quality on the lowest published level, CN 6-digit level, this will be done in the following section.

Table 16
Differences in million SEK and as a per cent of the edited value for all SITC1 codes for Method 5 for the first six months of 2004. Arrivals

SITC1	Measure	January	February	March	April	May	June
0	Difference in million SEK (imputed value – manual corrected value)	0.00	0.40	0.28	0.00	-0.01	-0.01
0	Total value when manually corrected (million SEK)	2,230	2,410	2,890	2,497	2,431	2,741
0	Difference in per cent (imputed value –manual corrected value)	0.00	0.02	0.01	0.00	0.00	0.00
1	Difference in million SEK (imputed value – manual corrected value)	0.00	-0.15	-0.25	0.00	0.00	0.00
1	Total value when manually corrected (million SEK)	407	416	534	555	547	613
1	Difference in per cent (imputed value –manual corrected value)	0.00	-0.04	-0.05	0.00	0.00	0.00
2	Difference in million SEK (imputed value – manual corrected value)	0.34	0.72	-0.20	0.79	0.11	0.83
2	Total value when manually corrected (million SEK)	896	846	1,172	1,220	1,204	1,194
2	Difference in per cent (imputed value –manual corrected value)	0.04	0.08	-0.02	0.06	0.01	0.07
3	Difference in million SEK (imputed value – manual corrected value)	0.00	0.00	-0.01	0.00	0.00	0.00
3	Total value when manually corrected (million SEK)	1,590	1,458	2,841	2,203	3,194	2,337
3	Difference in per cent (imputed value –manual corrected value)	0.00	0.00	0.00	0.00	0.00	0.00
4	Difference in million SEK (imputed value – manual corrected value)	0.00	0.00	0.00	0.00	0.00	0.00
4	Total value when manually corrected (million SEK)	118	108	96	91	78	108
4	Difference in per cent (imputed value –manual corrected value)	0.00	0.00	0.00	0.00	0.00	0.00
5	Difference in million SEK (imputed value – manual corrected value)	0.93	0.76	0.00	0.61	-0.76	-0.40
5	Total value when manually corrected (million SEK)	4,723	4,917	5,346	5,102	4,882	5,301
5	Difference in per cent (imputed value –manual corrected value)	0.02	0.02	0.00	0.01	-0.02	-0.01
6	Difference in million SEK (imputed value – manual corrected value)	-0.86	-1.58	0.06	-1.22	1.37	-1.89
6	Total value when manually corrected (million SEK)	5,140	5,641	6,797	6,594	6,457	6,736
6	Difference in per cent (imputed value –manual corrected value)	-0.02	-0.03	0.00	-0.02	0.02	-0.03
7	Difference in million SEK (imputed value – manual corrected value)	0.08	-0.09	-0.92	1.92	-0.41	1.41
7	Total value when manually corrected (million SEK)	13,140	15,488	18,252	17,156	17,135	18,875
7	Difference in per cent (imputed value –manual corrected value)	0.00	0.00	-0.01	0.01	0.00	0.01
8	Difference in million SEK (imputed value – manual corrected value)	-0.55	0.32	1.02	-1.87	-0.53	-0.54
8	Total value when manually corrected (million SEK)	3,641	4,092	4,685	4,055	3,908	4,401
8	Difference in per cent (imputed value –manual corrected value)	-0.02	0.01	0.02	-0.05	-0.01	-0.01
9	Difference in million SEK (imputed value – manual corrected value)	0.00	0.00	0.00	0.00	0.00	0.00
9	Total value when manually corrected (million SEK)	9	6	13	6	8	5
9	Difference in per cent (imputed value –manual corrected value)	0.00	0.00	0.00	0.00	0.00	0.00

Table 17
Differences in million SEK and as a per cent of the edited value for all SITC1 codes for Method 5 for the first six months of 2004. Dispatches

SITC1	Measure	January	February	March	April	May	June
0	Difference in million SEK (imputed value – manual corrected value)	-0.01	0.00	1.07	0.00	0.04	0.00
0	Total value when manually corrected (million SEK)	1,206	1,198	1,432	1,403	1,242	1,436
0	Difference in per cent (imputed value –manual corrected value)	0.00	0.00	0.07	0.00	0.00	0.00
1	Difference in million SEK (imputed value – manual corrected value)	0.00	0.00	-0.40	0.00	0.00	0.00
1	Total value when manually corrected (million SEK)	61	68	71	104	106	106
1	Difference in per cent (imputed value –manual corrected value)	0.00	0.00	-0.57	0.00	0.00	0.00
2	Difference in million SEK (imputed value – manual corrected value)	0.00	0.00	0.02	0.00	0.00	0.00
2	Total value when manually corrected (million SEK)	2,406	2,521	3,122	2,882	2,751	3,101
2	Difference in per cent (imputed value –manual corrected value)	0.00	0.00	0.00	0.00	0.00	0.00
3	Difference in million SEK (imputed value – manual corrected value)	0.00	0.00	0.00	0.00	-0.66	0.00
3	Total value when manually corrected (million SEK)	1,380	1,694	1,858	1,813	2,088	2,242
3	Difference in per cent (imputed value –manual corrected value)	0.00	0.00	0.00	0.00	-0.03	0.00
4	Difference in million SEK (imputed value – manual corrected value)	0.00	0.00	0.00	-0.57	-0.53	0.00
4	Total value when manually corrected (million SEK)	74	73	81	78	114	112
4	Difference in per cent (imputed value –manual corrected value)	0.00	0.00	0.00	-0.73	-0.47	0.00
5	Difference in million SEK (imputed value – manual corrected value)	0.00	-0.14	0.04	0.46	0.00	2.26
5	Total value when manually corrected (million SEK)	4,094	4,117	4,231	4,703	3,959	5,139
5	Difference in per cent (imputed value –manual corrected value)	0.00	0.00	0.00	0.01	0.00	0.04
6	Difference in million SEK (imputed value – manual corrected value)	-0.73	-0.24	-0.71	-0.02	0.80	0.60
6	Total value when manually corrected (million SEK)	9,698	9,625	11,028	10,277	10,872	11,923
6	Difference in per cent (imputed value –manual corrected value)	-0.01	0.00	-0.01	0.00	0.01	0.01
7	Difference in million SEK (imputed value – manual corrected value)	-0.01	-0.01	-0.37	-1.33	-0.80	-0.78
7	Total value when manually corrected (million SEK)	12,702	15,484	18,669	16,664	18,887	20,134
7	Difference in per cent (imputed value –manual corrected value)	0.00	0.00	0.00	-0.01	0.00	0.00
8	Difference in million SEK (imputed value – manual corrected value)	0.43	0.00	-0.06	0.94	0.59	-1.57
8	Total value when manually corrected (million SEK)	2,763	3,145	3,611	3,292	3,113	3,538
8	Difference in per cent (imputed value –manual corrected value)	0.02	0.00	0.00	0.03	0.02	-0.04
9	Difference in million SEK (imputed value – manual corrected value)	0.00	0.00	0.00	0.00	0.00	0.00
9	Total value when manually corrected (million SEK)	179	153	168	110	125	178
9	Difference in per cent (imputed value –manual corrected value)	0.00	0.00	0.00	0.00	0.00	0.00

Evaluation on CN 6-digit level

Although it was discovered in the previous section that using method 5 does not generate large deviations on SITC 1-digit level it is still justified to also evaluate the effects on CN 6-digit level, which is the lowest level that Foreign Trade data is published in Sweden.

The evaluation on CN 6-digit level has been made by comparing differentces between imputed and manually edited values measured before unit price checking and then comparing the sum of these differences for each CN code with the total value for the code after unit price checking. Since evaluating on a disaggregated level like this involves a large number of codes the evaluation is made for a single month, June 2004. Furthermore, only the codes with the largest deviations are included in this report.

In Table 18 below, the codes with the largest differences in per cent are illustrated for dispatches. As can be seen from this table most of the differences are very small. The largest effect on the total is for CN 731822, where the absolute difference in per cent is 8.79. In comparison with the effect on the other codes, this may seem very high. When looking at what caused this large difference it was discovered that the edited value for this observation was in fact unusually low. When studying other reported values for the PSI it seems as if the manually edited value is erroneous and that the imputed value probably is more correct. Earlier in the report the problems regarding evaluation of proposed methods and which values to use as "true" values was discussed. This example is a good illustration of this problem, the manually corrected values are not always the "true" values.

When looking at the other codes, the deviations are a lot smaller. The second highest absolute difference in per cent is only 0.20, which must be considered to be a small difference. It can also be noted that only five codes have differrences that are noticeable, i.e. the difference in per cent being 0.01 or higher.

Table 18
The largest differences in per cent between current validation and method 5. June 2004. Values in SEK.Dispatches

CN6 code	Imputed value	Edited value	Difference	Absolute difference	Total value	Absolute difference in per cent of total value
731822	494,954	58	494,896	494,896	5,629,023	8.79
902480	231,127	242,384	-11,257	11,257	5,646,735	0.20
170191	113,319	113,515	-196	196	504,968	0.04
392690	2,055,257	2,037,902	17,355	17,363	180,839,860	0.01
521213	0	2	-2	2	32,739	0.01
842490	34,060	33,878	182	182	10,602,266	0.00
382200	631,987	632,773	-786	786	103,578,166	0.00
490110	99,158	99,146	12	12	2,111,774	0.00
490199	252,706	252,438	268	272	25,302,917	0.00
960810	13,425	13,448	-23	23	2,958,488	0.00

For positive differences the maximum value is 494,896 SEK, which is related to CN code 731822 discussed above. The median difference is only 97 SEK and the minimum difference is 3 SEK. For negative differences the minimum value is –11,257 SEK, which can be considered to be a very small difference

on a single CN 6-digit code. The median value is –1 SEK and the maximum value is 0.

The same studies have been made also for arrivals. In Table 19 below the CN codes with the largest differences in per cent are illustrated. The code with the highest differences has a difference of 2.88 per cent compared to the total value for the code. The second and third largest deviations are 2.68 and 1.33 per cent of the total value respectively. It can also be noted that for arrivals the ten codes with the largest difference have noticeable differences. Although there are noticeable differences between using method 5 and the edited value, the differences seem reasonable.

Table 19
The largest differences in per cent between current validation and method 5. June 2004. Arrivals

CN6 code	Imputed value	Edited value	Difference	Absolute difference	Total value	Absolute difference in per cent of total value
710100	52,166	57,638	-5,472	5,472	190,325	2.88
841239	252,232	221,750	30,482	30,482	1,138,763	2.68
842490	48,312	381,312	-333,000	333,000	25,137,647	1.33
811909	36,315	40,774	-4,459	4,459	853,211	0.52
871610	1,368,066	1,477,559	-109,493	109,493	29,527,703	0.37
761699	428,670	541,875	-113,205	113,205	40,334,037	0.28
940340	574,088	641,813	-67,725	81,345	37,056,332	0.22
842430	243,506	255,631	-12,125	12,125	16,089,771	0.08
848320	26,484	25,370	1,114	1,114	4,160,136	0.03
920930	14,588	14,632	-44	44	185,670	0.02

For the positive differences the maximum value is 30,482 SEK, which corresponds to the code with the second highest percentage difference in Table 19. The median value is 1,396 SEK and the minimum value is 1 SEK. For the negative differences the minimum value is -333,000 SEK, which corresponds to the code with the third highest difference in per cent in Table 19. The median value is –41 SEK and the maximum value is 0.

From the evaluations that have been made in this and previous sections it seems reasonable to propose method 5 to be used in the validation process. It would decrease the number of observations that are required to be manually validated, but it seems to still produce data of good quality.

Therefore this method is also evaluated by using the unit price checking process to establish that it will not create problems or an increase in manual work during this part of the production process. This evaluation is made in chapter 5.

4 The unit price checking

In this chapter the unit price checking process is discussed. Below is given a description of the current price checking process.

4.1 Description of the unit price checking process

After the validity checking period is over the unit price checking starts. It consists of a SAS application that produces lists of suspected lines. For each line that has been reported to Statistic Sweden during the last month the application computes a number of unit prices: price per kilo, price per supplementary unit and the ratio between weight and supplementary unit. These unit prices are compared to the unit prices that have been reported previously. The lines that deviate from what has been previously reported can be called *suspected lines*.

But not all the suspected lines are checked. The aim is that only the suspected lines that could have an important effect on the published figures should be checked in order to minimize the manual work. This idea is realized by a score function that incorporates both the notion of suspicion and effect:

$$\text{Score}=\text{suspicion}*(\text{effect})^{p} \quad (1)$$

The X lines with the highest scores are checked. The number X can be varied depending on available resources. Below is described how the suspicion and effect is calculated.

4.1.1 Suspicion

The unit prices of each reported line is compared to the unit prices that have been reported previously on "similar lines". There are several ways to group lines into groups of similar lines. Similar lines can be lines with the same commodity code on CN6 level, or lines with the same commodity code on CN8-level, or lines with the same commodity code on CN8-level and with the same PSI. The following levels of groupings are used:

1) Flow, CN8-code, PSI, previous 12 months, country (country for arrivals only)
2) Flow, CN8-code, PSI, previous 12 months
3) Flow, CN8-code, PSI
4) Flow, CN8-code
5) Flow, CN6-code
6) CN6-code

The application starts by trying to compare the line in question with similar lines with the same value on the variables according to point 1 above. If there exists at least n (at present n≈7) historical lines on this level of detail, the line is checked against the unit prices of these lines. If it does not exist n historical observations on this level the program continues by trying to compare the line to lines with the same values on the variables in point 2 above. This is done if there exists at least n observations on this level of detail and so on. On level 6 the program requires only a smaller number of similar lines because otherwise the line cannot be checked at all.

From the similar lines the lower and upper quartile of the unit prices are computed. The distance between the quartiles, the inter quartile distance, is also computed. If the unit price of the line in question is above the upper quartile the suspicion is computed as the distance from the upper quartile measured in number of inter quartile distances. In the same way, if the unit price of the line in question is below the lower quartile the suspicion is computed as the distance from the lower quartile measured in number of inter quartile distances. If the unit price of the line in question is between the quartiles the suspicion is zero.

A bit simplified[2] the "suspicion function" takes the following mathematical form:

$$\text{Suspicion} = \begin{cases} LQ-UP/(UQ-LQ) & \text{if } UP<LQ \\ UP-UQ/(UQ-LU) & \text{if } UP>UQ \end{cases} \quad (2)$$

where UP is the unit price (price per kilo, price per supplementary unit or kilo per supplementary unit), LQ is the lower quartile and UQ is the upper quartile.

The suspicion is calculated one time for each unit price, resulting in three measures of suspicion. Lets call them SuspicionPKG, SuspicionPSU and SuspicionKGSU. We have then chosen to calculate the overall measure for suspicion as 0.5 times the maximum of the three measures plus 0.5 times the mean of all three measures. I.e.;

$$\begin{aligned} \text{Suspicion} = {} & 0.5 * \max(\text{SuspicionPKG}, \text{SuspicionPSU}, \text{SuspicionKGSU}) + \\ & 0.5 * \text{mean}(\text{SuspicionPKG}, \text{SuspicionPSU}, \text{SuspicionKGSU}) \end{aligned} \quad (3)$$

4.1.2 Effect

The main thought is that the effect is calculated as the deviation of the value of the line from the expected value of the line, set in relation to the normal value of the study domain. This can be expressed as follows:

$$effect = \frac{|value - \exp ected.value|}{\sum_{study.domain} value} \quad (4)$$

The expected value of the line is the value we would have expected given all the information we have about the line. It can be computed in two ways:

$$\text{expected value} = \begin{cases} (\text{median of price per kg}) * (\text{weight of line}) \\ (\text{median of price per sup. unit}) * (\text{sup. unit of line}) \end{cases} \quad (5)$$

The medians of the price per kilo and price per supplementary unit is the median of the unit prices of the similar lines according to the most detailed possible grouping of the groupings 1 – 6 in section 4.1.1.

The formula (4) expresses the essence of our definition of the effect. However we have made some extensions. Given a level of aggregation (e.g. CN6) one might want to have better relative precision on larger codes than on smaller codes. A 5 % deviation on a large code might be considered

[2] The calculation is done on logged values since the distribution of unit prices is skewed. Also, for commodity codes with stable prices, the quartiles might coincide. To cope with that the term 0.1 is added in the denominator, which corresponds to 10 percent difference between the quartiles.

worse than a 5% deviation on a small code. In addition one might want better quality on a more aggregated aggregation level (e.g. total trade or SITC2) than on a more disaggregated (CN6 or CN8). Consider for example a CN8-code that normally has a published value of 1 million SEK and a CN4-code that also normally has a published value of 1 million SEK. One might want to correct the error that has effect on the CN4-code before one corrects the error that has effect on the CN8-code given that the errors are of the same magnitude in SEK.

These thoughts are incorporated into the expression for effect according to the following:

$$Effect = \frac{\left|Value - \exp ected.value\right|}{\sum_{study.domain} Value} \cdot \frac{1}{O_{\substack{level.of \\ aggregation}}} \cdot f^{\,{}^{10}\log\left(\sum_{\substack{level.of \\ aggregation}} Value\right)}, \quad (6)$$

where *O* is a factor which is determined for each level of aggregation (e.g. CN8, CN6 etc.). The factor indicates the importance of the study domain. The smaller the factor is, the more important the study domain is considered to be. The management according to their own judgement sets the factors. Management also sets the factor *f*. It should be somewhere between 1 and 10. A factor of 1 implies that 1% deviation from the correct value is equally harmful on a small code as it is on a large code. A factor of 10 on the other hand implies that a 1 000 SEK deviation is equally harmful on a small code as it is on a large code.

The levels of aggregation judged important are total arrival and total dispatches, SITC2, SITC3, CN6 and certain CN8 codes. These levels of aggregation each have a specified *O*-value. The effect is calculated for each of these levels of aggregation and the maximum effect is used.

As mentioned before the expected value can be computed by using the price per kilo as well as by using the price per supplementary unit. As a result we get two measures of the effect. The final measure of effect is a linear combination between the two.[3]

4.2 Methods

In this section we discuss the approaches that could be adopted in the unit price checking in order to increase the automation or in other way decrease the manual work.

The simplest way to decrease the manual work related to unit price checking would be to adopt the "do nothing"- approach. This would mean that the unit price checking would be removed from the production process and we would rely only on the output checking in finding errors.

A less dramatic method would be to decrease the number of checked lines without completely removing the unit price checking. This could easily be done since we use a selective editing method with a score function.

[3] 0.3 times the smallest of the two and 0.7 times the largest of the two.

The two methods above decrease the manual work in the unit price checking but are not methods of increased automation in the sense that no variable values are imputed. The unit price checking method consists of checks of the unit prices (price per kilo, price per supplementary unit etc). A suspected unit price, e.g. a suspected price per kilo, indicates that either the value or the weight (or both) is incorrect. To be able to automatically impute a better price per kilo one would have to know which variable to change. This dilemma is discussed in chapter 6, proposals for further studies.

Another approach could be to accept that it is difficult to know which variable value to change. One could instead replace all variable values, i.e. the weight, supplementary unit and value of the line could be replaced with the weight, supplementary unit and value of a previously reported and checked line. This could be called nearest neighbour imputation. Practically you could replace the line with one of the "similar lines" which has been used to check the line. This is also discussed in chapter 6, proposals for further study.

The first two methods are discussed in the next section.

4.3 Evaluation of the methods

The unit price checking is much more easy to evaluate than the validity checking. In the production process the original values are stored and the values obtained after checking can be obtained from the Intrastat table at any time. As a consequence we get the variable values of each variable before and after checking. Weights, supplementary units and values before checking can be summed by the commodity codes before checking and they can be compared to the weights, supplementary units and values after checking summed by the commodity codes after checking. The differences between the summed weights, supplementary units and values before and after checking can be computed.

4.3.1 Evaluation of method 1

The most simple way to decrease the manual work needed in the unit price checking is simply not to do any unit price checking. To get an estimation of the effect of the unit price checking the absolute difference between the edited and unedited value can be computed for each line. The sum of these absolute differences is shown in Table 20 for arrivals and in Table 21 for dispatches.

Table 20
Arrivals. Sum of absolute differences between unedited and edited value in million SEK

Reference month	Summed absolute differences between unedited and edited value (Million SEK)	Total edited value (Million SEK)	Per cent of total value
January	2,843	31,895	8.9
February	242	35,382	0.7
March	489	42,625	1.1
April	289	39,478	0.7
May	401	39,843	1.0
June	519	42,310	1.2
Average	797	38,589	2.1

Table 21
Dispatches. Sum of absolute differences between unedited and edited value in million SEK

Reference month	Summed absolute differences between unedited and edited value (Million SEK)	Total edited value (Million SEK)	Per cent of total value
January	103	34,564	0.3
February	249	38,077	0.7
March	220	44,272	0.5
April	106	41,326	0.3
May	1,140	43,259	2.6
June	555	47,908	0.2
Average	395	41,568	1.0

From the tables above it is seen that for arrivals the sum of the absolute differences in SEK[4] varies from 242 million in February to 2,843 million in January. Expressed as per cent of the total value the variation is from 0.7 % to 8.9 % of the total value for the month. For dispatches the variation in SEK is from 103 million to 1,140 million and from 0.3 % to 2.6 %. Thus, the effect of the unit price checking is larger on arrivals than on dispatches.

The effect on the total value of arrivals and dispatches can be seen in Table 22 and Table 23. It can be noted that the positive and negative differences on individual lines cancel each other out to some extent when we sum them together. That is why the values in Table 22 and Table 23 are smaller than the values in Table 20 and Table 21.

[4] 1 Euro ≈ 9 SEK.

Table 22
Sum of differences between unedited and edited value in million SEK. Arrivals

Reference month	Difference in total value (Million SEK)	Total edited value (Million SEK)	Per cent difference in total value
January	2,750	31,895	8.6
February	158	35,382	0.4
March	284	42,625	0.7
April	229	39,478	0.6
May	340	39,843	0.9
June	372	42,310	0.9
Average	689	38,589	1.8

Table 23
Sum of differences between unedited and edited values in million SEK. Dispatches

Reference month	Difference in total value (Million SEK)	Total edited value (Million SEK)	Per cent difference in total value
January	31	34,564	0.1
February	87	38,077	0.2
March	-61	44,272	-0.1
April	59	41,326	0.1
May	1,094	43,259	2.5
June	503	47,908	1.0
Average[5]	306	41,568	0.7

From the table it can be seen that the effect on total arrivals varies from 0.4 % in February to an extreme 8.6 % in January. In SEK the effect varies from 158 million to 2,750 million. For dispatches the effect varies from –0.1 % to 2.5 % and in SEK the effect varies from –61 million to 1,094 million.

In total 1,185 invoiced values have been changed the for the reference months of the first half-year 2004. There are more changes that make the edited value larger (negative differences) than there are changes that make the edited value smaller (positive differences) (476 compared to 392). However, the positive differences are generally much bigger changes. The median positive difference is 1.2 million SEK while the median negative difference is 0.2 million SEK. The mean positive difference is 15 million SEK while the mean negative difference is 1.4 million SEK. This is why the edited total value is almost always smaller than the unedited value. The only exception is for dispatches in March where the editing has made the total value larger.

Some lines are deleted due to editing. It is interesting to see whether they constitute a big fraction of the changes described above. The deleted lines are examined in Table 24 below.

[5] Of absolute differences.

Table 24
Number of deleted lines and the value of deleted lines in million SEK

Reference month	Arrivals		Dispatches	
	Number of deleted lines	Value of deleted lines	Number of deleted lines	Value of deleted lines
January	26	47	31	22
February	47	62	33	10
March	25	29	8	29
April	14	19	20	21
May	32	136	31	90
June	34	72	16	14
Average	30	61	23	31

As can be seen, lines corresponding to a value of on the average 61 million SEK are deleted each month for arrivals and 31 million SEK for dispatches. The deleted lines may sometimes be because a whole report was deleted due to errors. A new report is then sent in so the effect on total value might be smaller than indicated in Table 24.

The effects can also be evaluated by commodity. In Table 25 and Table 26 the effects on SITC1 is presented for arrivals and dispatches respectively. For arrivals, out of the 60 computed differences, 18 were larger than 1 % of the edited value. The largest difference in per cent was for the SITC1 code number 7 where 19.4 % difference was obtained in January. Out of the 60 differences in SEK, 10 were larger than 50 million and the highest was on SITC code number 7 where the difference was 2,549 million.

For dispatches, six of the differences in per cent were larger than 1 %. The largest difference was on SITC1 code number 8 in May, where a 26.7 % difference was observed. Out of the 60 differences in SEK, Nine were larger than 50 million and the largest difference was for SITC1 code number 8 where an 830 million difference was recorded in May.

Table 25
Differences in million SEK and as per cent of the edited value for all SITC1 codes for the first six months of 2004. Arrivals

SITC1	Measure	January	February	March	April	May	June
0	Difference between unedited and edited value (million SEK)	26	42	19	20	4	-4
0	Total value when manually edited (million SEK)	2,230	2,410	2,890	2,497	2,431	2,741
0	Per cent difference between unedited and edited value	1.2	1.7	0.7	0.8	0.1	-0.2
1	Difference between unedited and edited value (million SEK)	-1	1	0	0	6	14
1	Total value when manually edited (million SEK)	407	416	534	555	547	613
1	Per cent difference between unedited and edited value	-0.2	0.2	0.0	0.0	1.1	2.2
2	Difference between unedited and edited value (million SEK)	9	22	11	-1	22	0
2	Total value when manually edited (million SEK)	896	846	1,172	1,220	1,204	1,194
2	Per cent difference between unedited and edited value	1.0	2.6	0.9	-0.1	1.9	0.0
3	Difference between unedited and edited value (million SEK)	0	-7	0	0	0	0
3	Total value when manually edited (million SEK)	1,590	1,458	2,841	2,203	3,194	2,337
3	Per cent difference between unedited and edited value	0.0	-0.5	0.0	0.0	0.0	0.0
4	Difference between unedited and edited value (million SEK)	0	4	0	0	-1	0
4	Total value when manually edited (million SEK)	118	108	96	91	78	108
4	Per cent difference between unedited and edited value	0.0	3.8	0.0	0.0	-1.4	0.0
5	Difference between unedited and edited value (million SEK)	30	12	11	40	202	9
5	Total value when manually edited (million SEK)	4,723	4,917	5,346	5,102	4,882	5,301
5	Per cent difference between unedited and edited value	0.6	0.2	0.2	0.8	4.1	0.2
6	Difference between unedited and edited value (million SEK)	15	17	97	82	8	35
6	Total value when manually edited (million SEK)	5,140	5,641	6,797	6,594	6,457	6,736
6	Per cent difference between unedited and edited value	0.3	0.3	1.4	1.2	0.1	0.5
7	Difference between unedited and edited value (million SEK)	2,549	68	14	86	93	314
7	Total value when manually edited (million SEK)	13,140	15,488	18,252	17,156	17,135	18,875
7	Per cent difference between unedited and edited value	19.4	0.4	0.1	0.5	0.5	1.7
8	Difference between unedited and edited value (million SEK)	121	-1	132	3	5	4
8	Total value when manually edited (million SEK)	3,641	4,092	4,685	4,055	3,908	4,401
8	Per cent difference between unedited and edited value	3.3	0.0	2.8	0.1	0.1	0.1
9	Difference between unedited and edited value (million SEK)	0	0	0	0	0	0
9	Total value when manually edited (million SEK)	9	6	13	6	8	5
9	Per cent difference between unedited and edited value	2.7	4.2	0.7	0.4	0.0	0.0

Table 26
Differences in million SEK and as a per cent of the edited value for all SITC1 codes for the first six months of 2004. Dispatches

SITC1	Measure	January	February	March	April	May	June
0	Difference between unedited and edited value (million SEK)	-2	2	0	-1	3	-2
0	Total value when manually edited (million SEK)	1,206	1,198	1,432	1,403	1,242	1,436
0	Per cent difference between unedited and edited value	-0.2	0.1	0.0	-0.1	0.2	-0.2
1	Difference between unedited and edited value (million SEK)	0	0	0	0	0	0
1	Total value when manually edited (million SEK)	61	68	71	104	106	106
1	Per cent difference between unedited and edited value	0.0	0.0	0.0	0.0	-0.3	0.0
2	Difference between unedited and edited value (million SEK)	7	6	-31	0	7	-1
2	Total value when manually edited (million SEK)	2,406	2,521	3,122	2,882	2,751	3,101
2	Per cent difference between unedited and edited value	0.3	0.2	-1.0	0.0	0.3	0.0
3	Difference between unedited and edited value (million SEK)	0	7	-99	-3	0	0
3	Total value when manually edited (million SEK)	1,380	1,694	1,858	1,813	2,088	2,242
3	Per cent difference between unedited and edited value	0.0	0.4	-5.3	-0.2	0.0	0.0
4	Difference between unedited and edited value (million SEK)	0	0	0	0	0	0
4	Total value when manually edited (million SEK)	74	73	81	78	114	112
4	Per cent difference between unedited and edited value	0.0	0.1	0.0	0.2	0.0	-0.1
5	Difference between unedited and edited value (million SEK)	20	32	1	-2	89	35
5	Total value when manually edited (million SEK)	4,094	4,117	4,231	4,703	3,959	5,139
5	Per cent difference between unedited and edited value	0.5	0.8	0.0	0.0	2.2	0.7
6	Difference between unedited and edited value (million SEK)	-13	24	-3	26	80	104
6	Total value when manually edited (million SEK)	9,698	9,625	11,028	10,277	10,872	11,923
6	Per cent difference between unedited and edited value	-0.1	0.2	0.0	0.3	0.7	0.9
7	Difference between unedited and edited value (million SEK)	20	-6	22	41	86	97
7	Total value when manually edited (million SEK)	12,702	15,484	18,669	16,664	18,887	20,134
7	Per cent difference between unedited and edited value	0.2	0.0	0.1	0.2	0.5	0.5
8	Difference between unedited and edited value (million SEK)	-1	22	49	-1	830	270
8	Total value when manually edited (million SEK)	2,763	3,145	3,611	3,292	3,113	3,538
8	Per cent difference between unedited and edited value	0.0	0.7	1.4	0.0	26.7	7.6
9	Difference between unedited and edited value (million SEK)	0	0	0	0	0	0
9	Total value when manually edited (million SEK)	179	153	168	110	125	178
9	Per cent difference between unedited and edited value	0.0	0.0	0.0	0.0	0.0	0.0

It would be desirable to get an estimate of the effect of the unit price checking on a more detailed commodity code level, e.g. the CN6 level. The commodity codes on CN6 level are too many to be listed in the same way as for the SITC1 codes. The effect will have to be described by using tables of subsets of the commodity codes. We also limit our analysis to a single reference month, June 2004.

For arrivals in June 2004, 165 CN6 codes had a changed total value due to editing. Of these differences 88 were positive and 77 were negative. The median positive difference is 338,000 SEK and the median negative change is –173,000 SEK. The largest positive difference on a single CN6 code is about 141 million SEK and the largest negative difference is about –19 million SEK.

For dispatches in June 2004, 111 CN6 codes had a changed total value due to editing, of these 61 were positive and 50 were negative. The median positive difference is 147,000 SEK and the median negative difference is –130,000 SEK. The largest positive difference is about 266 million and the largest negative difference is about - 8 million SEK.

The ten largest absolute differences for arrivals and dispatches are shown in Table 27 and Table 28 below. A large absolute difference combined with a large difference expressed as a per cent is very harmful. In that perspective e.g. the CN6 code 852910 on arrivals and e.g. the CN6 codes 902150 and 340212 on dispatches has benefited a lot from the editing

Table 27
The largest absolute differences in thousand SEK. Arrivals

CN6 code	Unedited value	Edited value	Difference	Difference in per cent of edited value
851790	794,256	653,322	140,933	22
852910	272,376	145,147	127,228	88
860800	13,955	33,311	-19,356	-58
761699	54,079	40,334	13,745	34
841370	68,231	54,945	13,285	24
852520	663,381	650,504	12,877	2
841899	48,509	37,143	11,365	31
330499	70,349	60,394	9,955	16
842481	17,671	8,499	9,172	108
853890	87,189	78,098	9,091	12

Table 28
The largest absolute differences in thousand SEK. Dispatches

CN6 code	Unedited value	Edited value	Difference	Difference in per cent of edited value
902150	374,372	108,004	266,368	247
870323	1 297,728	1,256,703	41,025	3
481840	174,903	134,769	40,134	30
721922	124,431	90,689	33,742	37
340212	38,264	7,480	30,784	412
846789	83,147	58,854	24,293	41
720837	52,107	30,207	21,900	72
844360	39,938	26,888	13,050	49
870210	43,915	31,315	12,600	40
820412	16,151	7,667	8,484	111

Large differences in SEK are less harmful if they occur on CN6 codes with large values. On the other hand medium or even small sized differences can be quite harmful if they occur on small CN6 codes. This explains the necessity to investigate the distribution of the differences expressed as a per cent of the edited value. For arrivals, the maximum positive difference in per cent is 125,452 % and the median positive difference is 4.13 %. The maximum negative difference is –100 %, which corresponds to a CN6 code with no value before the editing but with a value after editing. This can happen when a commodity code is changed into another on one or more lines. The results for dispatches are similar.

Table 29 to Table 32 present the ten largest differences expressed as per cent of the edited value for arrivals and dispatches respectively. For each flow two tables are presented, one for the largest positive differences and one for the largest negative differences. For arrivals it can be noted e.g. the CN6 code 850240 which had a value of 8.6 million SEK before editing and was changed to 86,000 SEK after editing or the CN6 code 310560 which had a value of 240,000 SEK before editing but was changed to 8.1 million. Thus these effects are large both in SEK and per cent.

Table 29
The largest positive differences expressed as per cent of the edited value. Values in thousand SEK. Arrivals

CN6 code	Unedited value	Edited value	Difference	Difference in per cent of edited value
282490	180	0	179	125,452
300120	2,129	11	2,118	19,447
850240	8,637	86	8,551	9,958
270760	340	14	327	2,361
450310	537	25	512	2,061
854071	49	7	41	557
850213	2,127	459	1,669	364
730791	9,394	2,919	6,475	222
551422	4,764	1,958	2,806	143
440831	265	115	150	130

Table 30
The largest negative differences expressed as per cent of the edited value. Values in thousand SEK. Arrivals

CN6 code	Unedited value	Edited value	Difference	Difference in per cent of edited value
310260	0	106	-106	-100
470692	0	10	-10	-100
310560	240	8,056	-7,816	-97
230240	23	212	-189	-89
40520	96	391	-295	-76
440910	225	563	-338	-60
860800	13,955	33,311	-19,356	-58
390910	2,214	4,959	-2,745	-55
481110	2,309	3,839	-1,530	-40
321000	4,203	6,888	-2,685	-39

For dispatches the CN6 code 902150 can be noted. It originally had 374 million SEK but was changed to 108 million. The CN6 code 020230 was doubled from 2 million SEK to 4 million SEK.

Table 31
The largest positive differences expressed as per cent of the edited value. Values in thousand SEK. Dispatches

CN6 code	Unedited value	Edited value	Difference	Difference in per cent of edited value
390610	5,687	367	5,320	1,448
840731	101	7	94	1,392
340212	38,264	7,480	30,784	412
410449	1	0	1	397
902150	374,372	108,004	266,368	247
830220	6,267	2,386	3,881	163
902219	4,118	1,618	2,500	154
300120	5	2	3	132
846239	533	233	300	129
381800	43	20	23	114

Table 32
The largest negative differences expressed as a per cent of the edited value. Values in thousand SEK. Dispatches

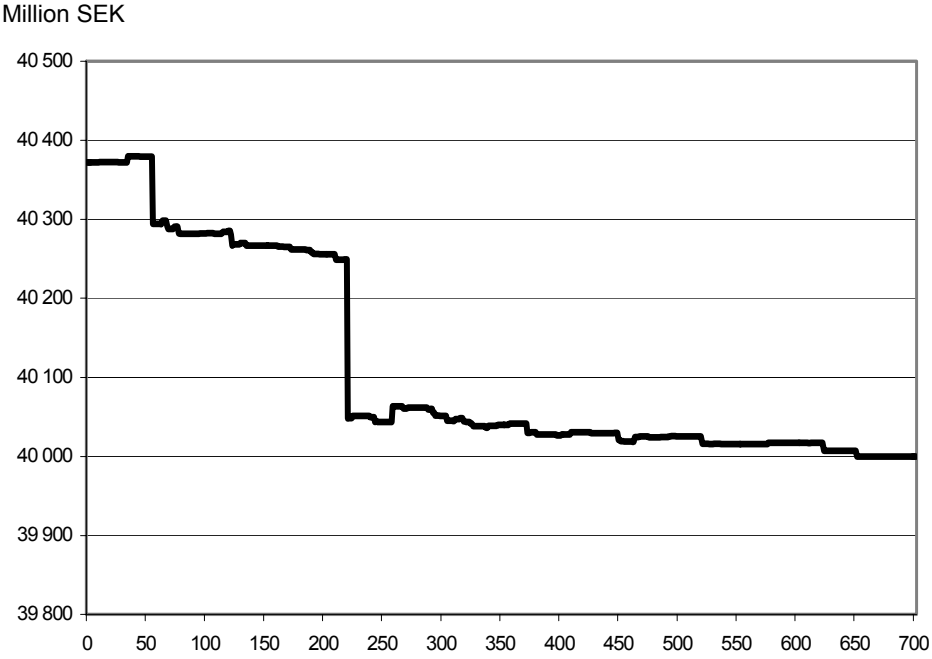
CN6 code	Unedited value	Edited value	Difference	Difference in per cent of edited value
281810	0	23	-23	-100
720441	0	82	-82	-100
846231	0	300	-300	-100
846820	68	718	-650	-91
020230	2,049	4,076	-2,027	-50
121220	8	12	-4	-30
040640	380	536	-155	-29
470311	950	1,226	-276	-23
440121	420	518	-98	-19
380630	18,601	22,597	-3,996	-18

Some CN6 codes had a value before editing but the value was removed due to editing. Three codes were deleted for arrivals and four codes were deleted for dispatches. The values of these codes are quite small both for arrivals and dispatches. For arrivals the largest unedited value is 79,000 SEK and for dispatches the largest unedited value is 371,000.

4.3.2 Evaluation of method 2

The number of lines checked manually could also be decreased without completely removing the unit price checking. Since we use a selective editing method with a score function this could easily be done. In Figure 2 below can be seen the effect on total arrivals in value when a certain number of lines is checked. We have used data for June 2004 where about 1500 lines were edited, 700 of these were reported for arrivals.

Figure 2
The estimated total value for arrivals in June 2004 as a function of the number of edited lines

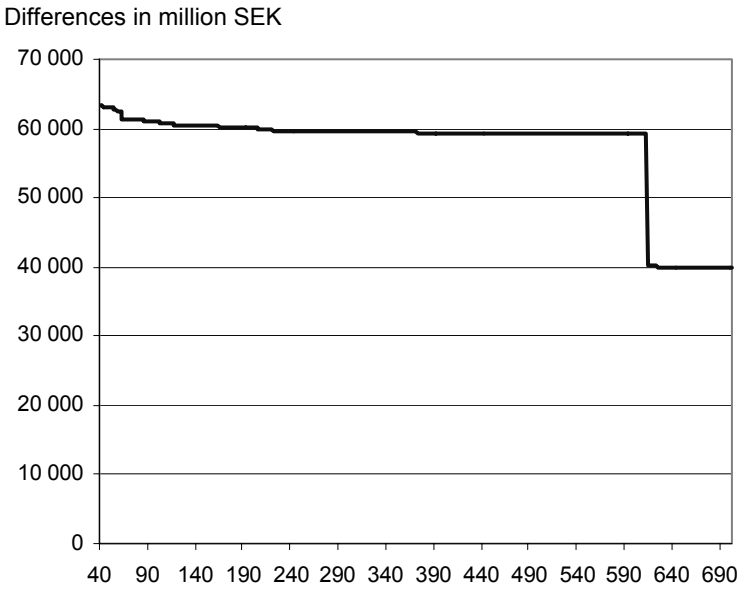


It is also interesting to examine the hit ratio, i.e. the number of changed lines in relation to the number of flagged lines. A hit means that at least one of the variables country code, commodity code, weight, supplementary unit or value is changed due to editing. The overall hit ratio is about 60 %. The hit ratio can also be computed by variable. The variable weight is the variable that is most often changed. The weight is changed on about 40 % of all flagged lines. The second most changed variable is supplementary unit, which is changed on about 15 % of all lines. This is very high since only about 20 % of all reported lines require a supplementary unit. Third comes the variable invoived value, which is changed in 10 – 15 % of all flagged lines, fourth is commodity code which is changed in about 7 % of all flagged lines and finally country code, which is changed in less than 5 % of the flagged lines.

As described above many of the changes are made on other variables than on the invoiced value, mostly on weight and supplementary unit. Invoiced value might be considered the most important variable and that is one of the reasons we have focused on evaluating the effect on value. Another reason is that it is the simplest variable to evaluate. The effect on aggregated weights on different aggregations is less easy to interpret. Normal variation in the weights of timber for example might completely hide large errors in the weight of jewellery for example. The supplementary unit is even more difficult to evaluate since the units of measurement differ between different commodity codes. To sum square meters and pieces is not relevant. A way out of this dilemma is to transform differences in weight and supplementary unit into differences in value. A difference in weight can be transformed into a difference in value by using the price per kilo of the edited line. For example if the difference in weight is 10 kg and the price per kilo after editing is 100 SEK per kg than the 10 kg difference corresponds to a 1,000

SEK difference. The differences originating from weight, supplementary unit and value can be summed for each line and that is the total difference for the line. The differences as a function of the number of edited lines can then be computed in the same way as in Figure 2. This is done in Figure 3. The first edited lines have such large differences that we exclude them from the figure. After about 600 edited lines for arrivals can be seen a large difference in the figure. This appears on the commoity code 8471 70 40 (memory units for computers) where a supplementary unit of over 87,000 was given but was then changed to 4. This difference of almost 87,000 units becomes 19,000 million SEK when multiplied by the price per supplementary unit of 218,000 SEK per unit.

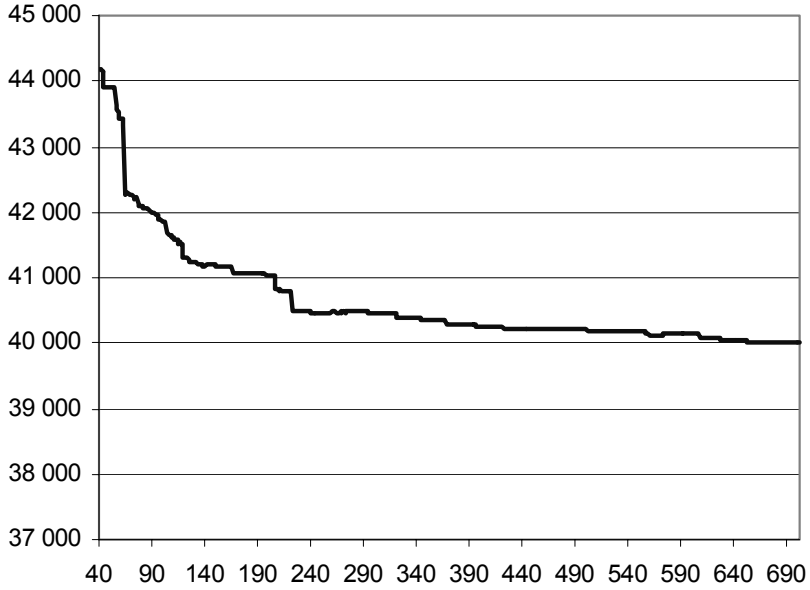
Figure 3
The impact on published statistics for arrivals in June 2004 as a function of the number of edited lines



If we remove this observation to get a better view of the other observations we get the figure below. The large difference in value in Figure 2 of 200 million SEK after about 220 lines can now be distinguished but now it is not that large in relation to other differences.

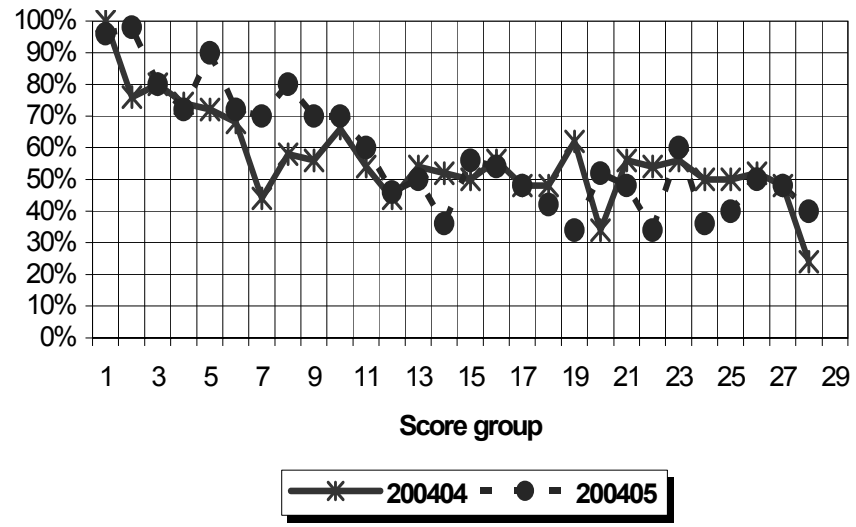
Figure 4
The impact on published statistics for arrivals in June 2004 as a function of the number of edited lines. One outlier removed

Differences in million SEK



The hit ratio can be computed in groups of e.g. 50 observations. This is done in Figure 5. The first group is created from the 50 observations with the highest score. The second group contains observations with scores that are ordered from 50 to 100 and so on. A hit means that at least one of the variables country code, commodity code, weight, supplementary unit or invoiced value is changed due to editing. In the figure data from April and May 2004 are used. As can be seen the hit-ratio for the first score group is almost 100 %. The hit-ratio then gradually decreases and passes 50 % somewhere between score group 12 and score group 26, which corresponds to 600 and 1,300 checked lines respectively.

Figure 5
Hit-ratio as a function of the Score. Hit-ratio computed in groups of 50 observations



In this section we will test what would be the effect of a decrease in the number of edited lines by almost 700 lines. The data set resulting from the normal editing process, where 1,500 lines are edited, will be compared to a data set that would be the result if about 800 lines were edited. To simplify notation we call these data sets the fully edited data set and the partially edited data set.

The sums of the absolute differences between the partially edited data set and the fully edited data set can be seen in the tables below. When comparing these tables to the tables for method 1 one can see that most of the errors are found on the first 800 lines. When no lines were edited we missed on average 797 million SEK of errors on arrivals and 395 million SEK on dispatches.[6] If 800 lines were to be edited instead of the 1,500 lines edited today, on average 80 million SEK of errors would be missed on arrivals and on average 54 million SEK on dispatches.

[6] These averages are affected by two large observations, the value for January for arrivals and the value for May for dispatches. When the larges and smallest observation are removed the averages becomes 425 million SEK for arrivals and 283 million SEK for dispatches.

Table 33
Arrivals. Sum of absolute differences between partially edited data set and fully edited data set in million SEK

Reference month	Summed absolute differences (Million SEK)	Total value when fully checked (Million SEK)	Per cent of total value
January	66	31,895	0.21
February	39	35,382	0.11
March	114	42,625	0.27
April	90	39,478	0.23
May	61	39,843	0.15
June	112	42,310	0.26
Average	80	38,589	0.21

Table 34
Dispatches. Sum of absolute differences between partially edited data set and fully edited data set in million SEK

Reference month	Summed absolute differences (Million SEK)	Total value when fully checked (Million SEK)	Per cent of total value
January	31	34,564	0.09
February	53	38,077	0.14
March	26	44,272	0.06
April	25	41,326	0.06
May	49	43,259	0.11
June	142	47,908	0.30
Average	54	41,568	0.13

When examining the effect on total arrivals and total dispatches the same pattern emerges. The effect on arrivals drops from on the average 689 million SEK[7] to on the average 42 million SEK. The effect on dispatches drops from on average 285 million SEK[8] to 24 million SEK.

Table 35
Sum of differences between partially edited data set and fully edited data set in million SEK. Arrivals

Reference month	Difference in total value (Million SEK)	Total value when fully edited (Million SEK)	Per cent difference in total value
January	47	31,895	0.15
February	16	35,382	0.05
March	34	42,625	0.08
April	56	39,478	0.14
May	38	39,843	0.09
June	61	42,310	0.15
Average	42	38,589	0.11

[7] 306 milllion SEK when the largest and smallest observations are removed.

[8] 178 milllion SEK when the largest and smallest observations are removed.

Table 36
Sum of differences between partially edited data set and fully edited data set in million SEK. Dispatches

Reference month	Difference in total value (Million SEK)	Total value when fully edited (Million SEK)	Per cent difference in total value
January	-4	34,564	-0.01
February	21	38,077	0.05
March	9	44,272	0.02
April	1	41,326	0.00
May	19	43,259	0.04
June	99	47,908	0.21
Average	24	41,568	0.06

When examining the effect on SITC1 one can see that for arrivals the number of differences in per cent exceeding 1.0 % drops from 18 to only 3 when the number of edited lines is increased from zero to 800. The number of differences attaining or exceeding 0.5 % is 4. The largest difference in per cent was 2.3 %. None of the differences in SEK now exceeds 50 million and 8 exceed 10 million SEK. The largest difference is 40 million SEK.

For dispatches the number of differences exceeding 1.0 % has dropped from 6 to 0. The number of differences attaining or exceeding 0.5 % is only 2. The largest difference in per cent was 0,7 %. The number of differences exceeding 50 million SEK has dropped from 9 to 1 and the number of differences exceeding 10 million SEK is 4. The largest difference was 55 million SEK.

Table 37
Differences between partially edited and fully edited data set in million SEK and as per cent of the edited value for all SITC1 codes for the first six months of 2004. Arrivals

SITC1	Measure	January	February	March	April	May	June
0	Difference between unedited and edited value (million SEK)	0	4	6	0	0	-4
0	Total value when manually edited (million SEK)	2,230	2,410	2,890	2,497	2,431	2,741
0	Per cent difference between unedited and edited value	0.0	0.2	0.2	0.0	0.0	-0.2
1	Difference between unedited and edited value (million SEK)	-1	1	0	0	7	0
1	Total value when manually edited (million SEK)	407	416	534	555	547	613
1	Per cent difference between unedited and edited value	-0.2	0.2	0.0	0.0	1.2	0.1
2	Difference between unedited and edited value (million SEK)	3	-1	27	-1	0	1
2	Total value when manually edited (million SEK)	896	846	1,172	1,220	1,204	1,194
2	Per cent difference between unedited and edited value	0.3	-0.1	2.3	0.0	0.0	0.1
3	Difference between unedited and edited value (million SEK)	0	1	0	0	0	0
3	Total value when manually edited (million SEK)	1,590	1,458	2,841	2,203	3,194	2,337
3	Per cent difference between unedited and edited value	0.0	0.1	0.0	0.0	0.0	0.0
4	Difference between unedited and edited value (million SEK)	0	0	0	0	-1	0
4	Total value when manually edited (million SEK)	118	108	96	91	78	108
4	Per cent difference between unedited and edited value	0.0	0.0	0.0	0.0	-1.4	0.0
5	Difference between unedited and edited value (million SEK)	8	0	0	28	10	16
5	Total value when manually edited (million SEK)	4,723	4,917	5,346	5,102	4,882	5,301
5	Per cent difference between unedited and edited value	0.2	0.0	0.0	0.5	0.2	0.3
6	Difference between unedited and edited value (million SEK)	-3	6	0	14	4	2
6	Total value when manually edited (million SEK)	5,140	5,641	6,797	6,594	6,457	6,736
6	Per cent difference between unedited and edited value	-0.1	0.1	0.0	0.2	0.1	0.0
7	Difference between unedited and edited value (million SEK)	36	9	-3	14	16	40
7	Total value when manually edited (million SEK)	13,140	15,488	18,252	17,156	17,135	18,875
7	Per cent difference between unedited and edited value	0.3	0.1	0.0	0.1	0.1	0.2
8	Difference between unedited and edited value (million SEK)	3	-3	5	0	1	7
8	Total value when manually edited (million SEK)	3,641	4,092	4,685	4,055	3,908	4,401
8	Per cent difference between unedited and edited value	0.1	-0.1	0.1	0.0	0.0	0.2
9	Difference between unedited and edited value (million SEK)	0	0	0	0	0	0
9	Total value when manually edited (million SEK)	9	6	13	6	8	5
9	Per cent difference between unedited and edited value	0.0	0.0	0.0	0.0	0.0	0.0

Table 38
Differences between partially edited and fully edited data set in million SEK and as per cent of the edited value for all SITC1 codes for the first six months of 2004. Dispatches

SITC1	Measure	January	February	March	April	May	June
0	Difference between unedited and edited value (million SEK)	0	0	0	-1	3	0
0	Total value when manually edited (million SEK)	1,206	1,198	1,432	1,403	1,242	1,436
0	Per cent difference between unedited and edited value	0.0	0.0	0.0	-0.1	0.2	0.0
1	Difference between unedited and edited value (million SEK)	0	0	0	0	0	0
1	Total value when manually edited (million SEK)	61	68	71	104	106	106
1	Per cent difference between unedited and edited value	0.0	0.0	0.0	0.0	-0.3	0.0
2	Difference between unedited and edited value (million SEK)	-1	-4	0	0	0	-2
2	Total value when manually edited (million SEK)	2,406	2,521	3,122	2,882	2,751	3,101
2	Per cent difference between unedited and edited value	0.0	-0.2	0.0	0.0	0.0	-0.1
3	Difference between unedited and edited value (million SEK)	0	0	0	0	0	0
3	Total value when manually edited (million SEK)	1,380	1,694	1,858	1,813	2,088	2,242
3	Per cent difference between unedited and edited value	0.0	0.0	0.0	0.0	0.0	0.0
4	Difference between unedited and edited value (million SEK)	0	0	0	0	0	0
4	Total value when manually edited (million SEK)	74	73	81	78	114	112
4	Per cent difference between unedited and edited value	0.0	0.0	0.0	0.2	0.0	-0.1
5	Difference between unedited and edited value (million SEK)	4	0	1	0	20	-1
5	Total value when manually edited (million SEK)	4,094	4,117	4,231	4,703	3,959	5,139
5	Per cent difference between unedited and edited value	0.1	0.0	0.0	0.0	0.5	0.0
6	Difference between unedited and edited value (million SEK)	-8	-1	-4	8	-2	43
6	Total value when manually edited (million SEK)	9,698	9,625	11,028	10,277	10,872	11,923
6	Per cent difference between unedited and edited value	-0.1	0.0	0.0	0.1	0.0	0.4
7	Difference between unedited and edited value (million SEK)	3	5	-1	1	-1	55
7	Total value when manually edited (million SEK)	12,702	15,484	18,669	16,664	18,887	20,134
7	Per cent difference between unedited and edited value	0.0	0.0	0.0	0.0	0.0	0.3
8	Difference between unedited and edited value (million SEK)	-2	22	12	-7	-1	4
8	Total value when manually edited (million SEK)	2,763	3,145	3,611	3,292	3,113	3,538
8	Per cent difference between unedited and edited value	-0.1	0.7	0.3	-0.2	0.0	0.1
9	Difference between unedited and edited value (million SEK)	0	0	0	0	0	0
9	Total value when manually edited (million SEK)	179	153	168	110	125	178
9	Per cent difference between unedited and edited value	0.0	0.0	0.0	0.0	0.0	0.0

We now examine the amount of errors lost by CN6 code if 800 instead of 1,500 lines are edited. Differences are computed between the value resulting from an editing process including 800 lines and the value resulting from an editing process including 1,500 lines. These differences are computed by CN6 code. As before this is done for the reference month June 2004. For arrivals, the median positive difference is now 230,000 SEK (a decrease from 2.7 million) and the maximum positive difference is now about 66 million SEK. The median negative difference is –189,000 SEK and the minimum negative difference is about -66 million SEK. [9]

For dispatches the median positive difference is 132,000 SEK and the maximum positive difference is 40 million SEK. The median negative difference is –130,000 SEK and the minimum negative difference is –7.5 million SEK

In Table 39 below it can be seen the ten largest absolute differences for arrivals. The CN6 code 850240 also has a very large difference in per cent. By editing 800 instead of 1,500 lines the value would have been 8.6 million SEK instead of the correct 86,000 SEK.

Table 39
The largest absolute differences in thousand SEK. Arrivals

CN6 code	Partially edited value	Fully edited value	Difference	Difference in per cent of fully edited value
852990	467,015	532,943	-65,928	-12
852910	210,994	145,147	65,847	45
841899	48,509	37,143	11,365	31
330499	70,349	60,394	9,955	16
853890	87,189	78,098	9,091	12
850240	8,637	86	8,551	9,958
850440	97,314	105,865	-8,551	-8
392690	194,620	186,902	7,718	4
841790	20,023	12,384	7,639	62
841370	62,313	54,945	7,367	13

The ten largest absolute differences for dispatches are shown in Table 40. The CN6 code 820412 has a large difference in per cent and a fairly large difference in SEK.

[9] The negative difference of –66 million comes from the CN6 code 852990. It had a completely unedited value of 534 million SEK. The first 800 lines edited made the value decrease to 467 million (-67 million). Further editing from 800 to 1,500 lines made the value increase by 66 million to 533 million. 8 lines were involved in this process. Some lines originally had the CN6 code 852990 but it was changed into some other CN6 code thereby decreasing the value. Some lines had the same CN6 code before and after editing but the value was changed thereby either increasing or decreasing the value. Finally some lines originally had a different CN6 code but it was changed into 852990, thereby increasing the value.

Table 40
The largest absolute differences in thousand SEK. Dispatches

CN6 code	Partially edited value	Fully edited value	Difference	Difference in per cent of fully edited value
481840	174,903	134,769	40,134	30
846789	83,147	58,854	24,293	41
844360	39,938	26,888	13,050	49
870210	43,915	31,315	12,600	40
820412	16,151	7,667	8,484	111
441830	77,670	85,224	-7,553	-9
470321	888,150	894,180	-6,030	-1
392390	27,795	23,220	4,575	20
251710	18,623	14,516	4,107	28
380630	18,601	22,597	-3,996	-18

For arrivals the median positive difference is about 3 % and the largest positive difference is 9,957 %. The median negative difference is 2.5 % and the largest negative difference is 89 %. For dispatches the maximum positive difference is 1,392 % and the median positive difference is 17 %. The minimum negative difference is –100 %, which corresponds to a CN6 code with no value after 800 lines have been edited but got a value somewhere between 800 and 1,500 edited lines. The median negative difference is –1.4 %.

The ten largest positive differences are presented in Table 41.

Table 41
The largest positive differences expressed as per cent of the edited value. Values in thousand SEK. Arrivals

CN6 code	Partially edited value	Fully edited value	Difference	Difference in per cent of fully edited value
850240	8,637	86	8,551	9,958
854071	49	7	41	557
90220	58	27	30	111
340130	8,088	3,853	4,235	110
250410	1,994	1,078	916	85
841790	20,023	12,384	7,639	62
852910	210,994	145,147	65,847	45
842481	12,187	8,499	3,688	43
841899	48,509	37,143	11,365	31
950440	3,543	2,751	792	29

The ten largest negative differences are presented in Table 42.

Table 42
The largest negative differences expressed as per cent of the edited value. Values in thousand SEK. Arrivals

CN6 code	Partially edited value	Fully edited value	Difference	Difference in per cent of fully edited value
230240	23	212	-189	-89
40520	96	391	-295	-76
390910	2,214	4,959	-2,745	-55
611693	437	688	-251	-37
810199	177	246	-69	-28
843999	24,084	30,084	-6,000	-20
381129	7,599	9,075	-1,477	-16
852990	467,015	532,943	-65,928	-12
830170	2,511	2,792	-280	-10
480920	6,564	7,164	-600	-8

Table 43
The largest positive differences expressed as per cent of the edited value. Values in thousand SEK. Dispatches

CN6 code	Partially edited value	Fully edited value	Difference	Difference in per cent of fully edited value
840731	101	7	94	1,392
830220	6,267	2,386	3,881	163
902219	4,118	1,618	2,500	154
300120	5	2	3	132
381800	43	20	23	114
820412	16,151	7,667	8,484	111
320412	5,768	2,990	2,778	93
848041	843	508	334	66
844360	39,938	26,888	13,050	49
846789	83,147	58,854	24,293	41

Table 44
The largest negative differences expressed as per cent of the edited value. Values in thousand SEK. Dispatches

CN6 code	Partially edited value	Fully edited value	Difference	Difference in per cent of fully edited value
281810	0	23	-23	-100
40640	380	536	-155	-29
470311	950	1,226	-276	-23
380630	18,601	22,597	-3,996	-18
441830	77,670	85,224	-7,553	-9
711719	1,854	2,034	-180	-9
843330	647	691	-44	-6
480431	57,994	59,924	-1,930	-3
845129	1,010	1,035	-25	-2
902790	36,456	37,256	-800	-2

When editing 1,500 lines instead of 800 one CN6 code would disappear for arrivals and two CN6 codes would disappear for dispatches. The unedited value for the CN6 code disappearing for arrivals is 69,000 SEK. For dispatches the values for the two CN6 codes are 371,000 and 44,000 SEK.

5 The interaction between Validation and Unit Price checking

It has been discovered that sometimes variables that are imputed during the Validation process also have to be checked during the Price Checking process. These observations cannot be sent out to the PSIs in the ordinary mail-based system used in the Unit Price checking, because of the fact that the values that are indicated as possible errors are not the same figures as reported by the PSIs. This means that the clerks have to contact the PSI in order to get the correct values in the same way they would have done during the Validation process or manually impute the value. In both cases manual work is required.

This is why we devote this chapter to the interaction between the Validation and the unit price checking. In the first part of the chapter we describe the problem and its causes. Then we evaluate what the effect would be on the interaction if the thresholds used in the validation process were to be increased according to method 5 described previously. In the third part of the chapter we discuss what could be done to reduce the problem. It is established that we need to find a different method of imputation in the validation process. A test of a simplified version of such a method is performed and evaluated.

5.1 The problem and its causes

In Table 45 below can be seen the number of automatically imputed lines that were flagged in the unit price checking for the reference months January to June 2004.

Table 45
Total number of imputed observations flagged during Unit Price Checking in the actual production process

Reference month	Number of flagged observations
200401	80
200402	102
200403	119
200404	125
200405	124
200406	111
Total	680

A closer study for the flagged observations was made and it seems as if there are some codes where the problem is larger than others, for example 7113 19 00, Jewellery of precious metal. But the problem is after all not concentrated to few commodity codes. The problem appears over the whole range of codes.

There seems to be a number of reasons why automatically imputed lines are flagged in the unit price checking. One problem arises from the codes that require both weight and supplementary unit. If weight or supplemen-

tary unit is missing the automatic imputations are done from the value. In the unit price checking the weight per supplementary unit is compared to similar lines and the line is flagged.

Another example is when, for example, the supplementary unit is imputed from the value but the price per kilo is flagged in the unit price checking. Then the flagging doesn't have anything to do with the automatic imputation. These problems are probably difficult to decrease.

There can also be an error in the commodity code and as a result a commodity code is imputed automatically during the validation process. The imputed code may be different from the code intended by the PSI and the historical unit price for the imputed code may differ from the unit price for the intended code. As a result the line is flagged in the unit price checking.

Another problem with the price register is that the unit prices can stay at a level that is too high or too low. This is due to the fact that the recalculations of the price register, which is done almost every day, only includes values that do not deviate more than a specified percentage from the price calculated the previous time. The result of the current way of calculating the unit prices is that if the first lines introduced into the Intrastat system for a commodity code is in error the lines entered at a later period will not be included in the calculations, unless these are erroneous themselves.

A closer evaluative study of the causes of the problems can be done by performing the unit price checking on the data from our test database. The test data that is used is the data from the test where no changes were made compared to the ordinary production process.

In table 62 can be seen that when using the test data the number of flagged imputed observations is 718 for the first 6 months of 2004. Thus, in the actual production process fewer observations are flagged than in our test in the test database. This can be seen by comparing Table 45 and Table 46. The difference is 38 observations for the 6-month period. An explanation for this can be that in the actual production process both the imputation price register and the unit price checking prices change over time whereas those are fixed in our test. In our test we have also flagged the most important errors over the whole 6-month period whereas in the production process the most important errors for each month are flagged.

Table 46
Total number of imputed observations flagged during Unit Price Checking when tested in the test database

Reference month	Number of flagged observations
200401	99
200402	123
200403	132
200404	120
200405	116
200406	128
Total	718

The imputed lines can now be broken down by imputed variable. This is done in Table 47. Most flagged observations are imputed on supplementary unit. The second most flagged variable is commodity code with almost as many variable values flagged. Then come weight, country code and invoiced value in that order.

Table 47
Number of imputed variable values on observations flagged broken down by imputed variable

Variable	Number imputed variable values on observations flagged
Commodity code	320
Country code	18
Weight	115
Supplementary unit	344
Invoiced Value	4

The imputations of commodity codes can be broken down by type of imputation. This can be seen in table 64. As can be seen most of the flagged commodity codes are imputed from 4 digits. This is not because imputations from a 4-digit level are more common. Rather, when imputing from 4 digits there are a larger number of 8-digit codes to choose from than when imputing from 6 or 7 digits. That probably means that the unit prices vary more within a CN4 group than within a CN6 or CN7 group. If another code than the intended code is imputed the risk of a deviating unit price is larger when imputing from CN4 than when imputing from CN6 or CN7.

Table 48
Imputed commodity codes flagged in the unit price checking by type of imputation

Imputation type	Number of imputed variable values on flagged observations
Imputed from 4 digits	173
Imputed from 6 digits	75
Imputed from 7 digits	32
Old code changed to new	40

It can also be interesting to look at what type of control has flagged the imputed observations. The price per kilo is the control that flags most of the imputed observations but it is also the control that flags the most observations overall. The controls for weight per supplementary unit seems to stand for a large portion of the flagged imputed values.

Table 49
Type of control that has flagged the imputed observations

Control	Number of flagged observations
Price per kilo	245
Price per supplementary unit	128
Weight per supplementary unit (overall control)	185
Weight per supplementary unit (Control for certain supplementary units)	159

5.2 Unit price checking of data imputed using method 5

Earlier in the report we have evaluated methods to increase the number of imputations by increasing the thresholds for imputations. Here we test what effect the increase in thresholds used in method 5 will have on the unit price checking.

In Table 50 the number of imputed observations that would be flagged if 1,500 lines were edited in the unit price checking is presented. When comparing Table 50 with Table 46 one can see that the total number of flagged observations increase by 40 observations, from 718 to 758. On average this is an increase of 7 observations per month. It is also interesting to see what happens if method 5 is used together with a decrease in the number of checked lines in the unit price checking to 800 lines. This is also shown in Table 50 below.

Table 50
Total number of imputed observations flagged during Unit Price Checking when thresholds are increased moderately (method 5). 1500 lines edited

Reference month	Number of flagged observations, 1500 lines	Number of flagged observations, 800 lines
200401	108	64
200402	129	68
200403	135	77
200404	127	68
200405	121	79
200406	138	68
Total	758	424

If the number of checked lines is decreased by 47 % (from 1,500 to 800) the number of imputed flagged lines will decrease by 44 % (from 758 to 424). We find it a bit surprising that the number of flagged imputed lines is not decreased more.

We now discuss methods for reducing the problem of imputed variables being flagged in the unit price checking.

5.3 Methods for reducing the problem

The imputation in the validation process is done by commodity code and, if possible, also by country. There seem to be a large variation in the prices

between different countries. The unit price checking is however not primarily done by Country Code, rather the historical observations are grouped by PSIs as described in chapter 4. This difference in the imputation method and in the method for unit price checking causes imputed values to be flagged.

In the press releases the statistics are presented by commodity codes and countries. It might seem natural then to impute using prices per commodity code and country. But worth mentioning is that preliminary studies have indicated that PSI can explain more of the variation in the unit prices than country. That would suggest that imputations with prices per commodity code and PSI is more reliable.

A solution to the problem of imputed values being flagged in the unit price checking might be to use the unit prices calculated in the unit price checking as a basis for the imputation price register. Ideally one would use the same stepwise procedure as in the unit price checking, i.e. the program would start by trying to impute using the unit price per flow, CN8 code, PSI, using data from previous 12 months and country. If this was not possible because it was not enough observations to calculate the unit price the program would impute using the unit price per flow, CN8 code, PSI and using data from previous 12 months, i.e. regardless of country. If this were not possible either the program would use the next level of grouping and so on.

We have tested a very simple version of this procedure in our Intrastat test data system. We have taken the prices per commodity code calculated in the unit price checking (i.e. grouping number 4 in section 4.1.1) and replicated them for each country, i.e. all countries gets the same unit price for each commodity code.

In Table 51 below is shown the number of prices (regardless of country) in the current imputation price register and the number of prices (regardless of country) generated in the unit price checking. The number of prices generated in the unit price checking is smaller since one requires a certain number of observations to calculate the unit prices. In the table below seven observations were required. In the imputation price register this limit does not exist.

Table 51
Number of prices in Validation price register and in the price register in the Unit Price checking. All data regardless of country

	Price per kilo		Price per supplementary unit	
	Arrivals	Dispatches	Arrivals	Dispatches
Imputation price register	9,795	8,664	2,826	2,545
Price checking register	7,809	6,590	2,300	1,983
Difference	1,986	2,074	526	562

The system actually calculates prices per commodity code and country. In Table 52 below the number of prices actually calculated in the imputation price register is shown. As a comparison is given the number of prices that can be calculated from the unit price checking if the price per commodity code is simply replicated for each country. The last line in the table (the combined price register) shows the number of prices generated when the two price registers are combined into one single price register. This register is constructed by using the unit price checking prices as a base and adding prices from the current imputation price register when there is no price in the unit price register. An explanation for the larger number of prices generated for the unit price checking could be that in the unit price checking data for the ten new member states are included from Extrastat.

Table 52
Number of prices in the Validation price register and in the price register created from the prices in the Unit Price checking. Prices calculated by country

	Price per kilo		Price per supplementary unit	
	Arrivals	Dispatches	Arrivals	Dispatches
Imputation price register	79,187	83,753	24,401	24,371
Unit price checking register	203,034	171,340	59,800	51,558
Combined price register	210,866	179,808	62,175	54,113

We have used the combined, or alternative, price register to evaluate if it is possible to decrease the number of imputed values that are flagged in the unit price checking. Data for the first 6 months 2004 are run through the validation process in our test database. The resulting data is then run through the unit price checking. We first evaluate what effects the new price register has on the validation process.

5.3.1 Evaluation of the effect on the Validation process

The result from the test using the current price register is illustrated in Table 53 and the result from the test using the alternative price register is illustrated in Table 54. From the tables it is quite clear that the number of imputed observations is approximately the same. Furthermore, the differences in price registers do not change the value in SEK or weight to a large extent.

Table 53
Effect on the validation of using current price register. Period: 200401-200406

Variable	Deleted values			Imputed values		
	Number of observations	Value in SEK millions	Weight in tonnes	Number of observations	Value in SEK millions	Weight in tonnes
Total	1,462	2.10	2.11	47,969	3,943.37	390,384.82
Country Code	587	0.80	0.88	1,618	588.53	20,862.92
Commodity Code	913	1.36	1.28	24,839	2,080.04	300,476.79
Net Weight	20	0.01	0.02	8,892	191.83	7,700.14
Suppl. Unit	212	0.28	0.31	22,402	1,658.24	190,175.95
Invoiced Value	28	0.02	0.03	895	15.25	909.47

Table 54
Effect on the validation of using the price register from unit price checking. Period: 200401-200406

Variable	Deleted values			Imputed values		
	Number of observations	Value in SEK millions	Weight in tonnes	Number of observations	Value in SEK millions	Weight in tonnes
Total	1,454	2.08	2.11	48,037	3,955.32	391,406.15
Country Code	587	0.80	0.88	1,617	587.76	20,831.04
Commodity Code	912	1.35	1.29	24,880	2,084.37	300,996.77
Net Weight	21	0.01	0.02	8,876	191.93	7,694.86
Supplementary Unit	208	0.27	0.31	22,467	1,667.92	190,502.12
Invoiced Value	29	0.02	0.03	897	15.56	1,083.01

One of the reasons for evaluating this method already for the Validation Process is that the price register used for this method consists of a larger number of unit prices than the one currently used. As it turns out it does not change the validation process itself to a large extent. The method might still be justified if it decreases the amount of observations that has to be checked during the Price Checking Process.

5.3.2 Evaluation of the effect on the Unit Price checking

In Table 55 below can be seen the number of imputed observations that is flagged for each month when using the alternative price register. When comparing Table 46 and Table 55 can be seen that less imputed observations are flagged when the alternative price register is used then when the current price register is used. The difference is about 12 observations per month.

Table 55
Total number of imputed observations flagged during Unit Price Checking using the alternative price register

Reference month	Number of flagged observations
200401	95
200402	105
200403	121
200404	104
200405	108
200406	115
Total	648

The alternative price register should have the largest impact on the number of flagged weights, supplementary units and values. The number of flagged commodity codes and country codes cannot be affected by a change in price register. To improve the imputation of country and commodity code other measures are called for. In Table 56 below can be seen the number of flagged observations by imputed variable when using the current price register as well as when using the alternative price register. As expected the largest differences are for weight and supplementary unit.

Table 56
Number of imputed variable values on observations flagged. Data imputed using the current price register and data imputed using the alternative price register

Variable	Number imputed variable values on observations flagged		
	Current price register	Alternative price register	Difference
Commodity code	320	321	+1
Country code	18	18	0
Weight	115	91	-24
Supplementary unit	344	288	-56
Invoiced Value	4	3	-1

As before, the imputations of commodity codes can be broken down by type of imputation. This is done in Table 57 for data imputed using the current price register as well as for data imputed using the alternative price register. No large differences can be seen.

Table 57
Number of imputed commodity codes by type of imputation on observations flagged. Data imputed using the current price register and data imputed using the alternative price register

Imputation type	Number imputed variable values on observations flagged		
	Current price register	Alternative price register	Difference
Imputed from 4 digits	173	165	-8
Imputed from 6 digits	75	78	+3
Imputed from 7 digits	32	33	+1
Old code changed to new	40	45	+5

It can also be interesting to examine whether some types of controls have experienced larger changes than others. In table 75 can be seen that the largest change is for the overall control of weight per supplementary unit. Also the price per supplementary unit has experienced a decrease in the number of flagged observations.

Table 58
Type of control that has flagged the imputed observations

Control	Number of flagged observations		
	Current price register	Alternative price register	Difference
Price per kilo	245	239	-6
Price per supplementary unit	128	102	-26
Weight per supplementary unit (overall control)	185	153	-32
Weight per supplementary unit (Control for certain supplementary units)	159	153	-6

The result of the test seems promising. It is also likely that the results will be bigger if the method used for imputation and the method for unit price checking is further harmonized than what is done in this simple test. There is still potential for decreasing the number of flagged supplementary units and weights.

6 Proposals for further studies

During our work we have discovered aspects of the production process that can be improved and aspects that should be studied further. These are discussed in this section.

A lot of the manually corrected errors on commodity codes and country codes are due to missing codes. The imputation methods we use today cannot impute missing codes. Also many of the manually corrected errors on commodity code are due to the fact that the first 4-digits in the given code do not exist as a 4-digit commodity code. These kinds of errors are similar to the missing codes error in the sense that the given code cannot be used as a basis for imputation. A lot of incorrect country codes are also given which cannot be used for imputation.

These problems call for a method to impute missing codes as well as codes that in other ways cannot be imputed by the current methods regardless of how high the current thresholds for imputation are set. The procedure for deletion of observations is one such method. However this method is best for small value observations. For moderately large observations a different method might be better.

One approach could be to impute a commodity code or country code from the country code or commodity code of an observation that is very similar to the given observation. A similar observation could be from the same company or from a company within the same industry. Another approach could be to impute a random code from all possible codes. The probability of selection of each code could be proportional to the normal total value of the code. A drawback to these methods will probably be that the observations will be flagged in the unit price checking because the price per kilo implied by the observation might not match the price per kilo of the imputed commodity code.

In the current system used for imputation missing values, weights and supplementary units are imputed using the price per kilo and price per supplementary unit of the stated commodity codes. This process might be used in reverse. It may be possible to impute a commodity code given the price per kilo implied by the stated weight, supplementary unit and value. In that way the observations will at least not be flagged in the unit price checking.

In the current system, when commodity codes are imputed from a 4-, 6- or 7- digit level, the commodity code with the largest value is chosen. This might cause the unit price of the line to deviate from the other lines on the commodity code and the line will be flagged in the unit price checking. To avoid this the commodity code with the most compatible unit price could be selected out of the possible codes in the same CN4, CN6 or CN7 group.

Further harmonization of the validation process and the unit price checking is desirable to avoid that imputed weights, supplementary units and invoiced values are flagged in the unit price checking. The same methods should be used for imputation and unit price checking.

In order to be able to make changes to the imputation process and evaluate the effect resources should be allocated to a better documentation of the

imputation system. This should be done by IT personnel or in close collaboration with IT personnel.

In this report we have investigated if the manual work in the unit price checking can be decreased by decreasing the number of observations edited. No method of imputation has been proposed. A successful imputation would probably require that we could distinguish which variable is in error when e.g. the price per kilo deviates from the expected. It would be desirable to get an indication of whether it is the weight or the value that causes the price per kilo to deviate. Preliminary studies have shown that this can be done in certain cases. By calculating the suspicion for the weight and value separately using the same method as is used for the unit prices in the unit price checking we can get an indication of whether it is the weight or value that is most suspicious. The suspicion was calculated on each of the following levels of grouping if more than 4 observations are available in the group:

- flow, CN8, PSI, previous 24 months
- flow, CN8, PSI, previous 12 months
- flow, CN8, PSI, previous 12 months, country

The mean of the calculated suspicions are used as the measure for suspicion.

It is not clear whether the indications are strong enough to be used in a basis for deciding which variable to impute. Our hopes are not that high. The results might on the other hand be used in another context. If the view of Statistics Sweden is that invoiced value is a more important variable than weight and supplementary unit the results might be used in the unit price checking to flag expected errors in the invoiced value to a larger extent than expected errors in weight or supplementary unit. The suspicion of an error in the invoiced value might be incorporated into the suspicion function described in section 4.1.1. Preliminary studies have shown that the hit ratio for the variable invoiced value can be increased by this incorporation.

If it is not possible to distinguish which variable is causing a deviating unit price the imputation of a single variable is difficult. Another approach that we have started working on is to impute the whole observation rather than a single variable value. Our thought is that a similar, previously reported and edited line could replace a highly suspicious line. The similar line could be taken from one of the groupings described in section 4.1.1. This method would ensure that the resulting line is consistent, i.e. the imputation wouldn't create a line that is unreasonable in any way. If this method is adopted you should probably edit a proportion of the lines manually anyway. That would allow the unit prices to change over time.

6.1.1 Effect on respondent behaviour

At present Statistics Sweden contacts at least some of the respondents that makes mistakes in their reports. This makes the respondent aware that he or she has made a mistake and might also gives him or her an increased knowledge of the Intrastat system. Furthermore the respondent understands that it is important that the reports are correct. The result of all this may be that this particular mistake and other mistakes are less likely to be made in the future.

If fewer of the respondents are contacted, which would be the result of increased automated correction, this might lead to a change in the respondents' behaviour. After a while the respondents might think that it is not important to provide correct information and may also not notice that mistakes are committed.

The conclusion from the discussion in this section is that even if our investigation in this report should indicate that an increased automation has little effect on the published figures today, this might not be true in the future. The increased automated correction might lead to a change in the respondents' behaviour so that more mistakes are committed and the effect on published figures might be larger. This suggests that the effect of the increased automation must be investigated again in the future and that the level of automated correction may have to be adjusted.

Annex 1. Description of methods used for the validation process

In this annex a detailed description is made for the different methods that has been evaluated for the Validation process. The information is given in a table since the methods are based on changing the thresholds used for imputing or erasing observations. Table 59 describes the erroneous variable, the type of error and what kind of correction made for the different methods.

Table 59
Description of the thresholds used in Method 1 to 6. Value in SEK

Variable	Type of error	Correction	Threshold					
			Method 1	Method 2	Method 3	Method 4	Method 5	Method 6
Total invoiced amount	Incorrect sum	Imputed/converted	10[2]	10[2]	999[2]	10[2]	20[2]	10[2]
Country code	Not an EU country	Erased	0[1]	0[1]	0[1]	0[1]	0[1]	0[1]
Commodity code	Non-valid	Imputed from 4-digit level	500,000	100,000,000	100,000,000	500,000	1,000,000	500,000
Commodity code	Non-valid	Imputed from 6-digit level	1,000,000	100,000,000	100,000,000	1,000,000	2,000,000	1,000,000
Commodity code	Non-valid	Imputed from 7-digit level	1,000,000	100,000,000	100,000,000	1,000,000	2,000,000	1,000,000
Commodity code	Old code	Imputed from 4-digit level	500,000	100,000,000	100,000,000	500,000	1,000,000	500,000
Commodity code	Old code	Imputed from 6-digit level	1,000,000	100,000,000	100,000,000	1,000,000	2,000,000	1,000,000
Commodity code	Old code	Imputed from 6-digit level	1,000,000	100,000,000	100,000,000	1,000,000	2,000,000	1,000,000
Commodity code	Old code	Converted to a valid code	10,000,000	100,000,000	100,000,000	10,000,000	10,000,000	10,000,000
Net weight	Missing	Imputed/converted	250,000	100,000,000	100,000,000	250,000	500,000	250,000
Supplementary unit	Missing	Imputed/converted	10,000,000	100,000,000	100,000,000	10,000,000	10,000,000	10,000,000
Invoiced value	Missing	Imputed/converted	250,000	100,000,000	100,000,000	250,000	500,000	250,000
	General for all types of errors	Erased	6,000[3]	6,000[3]	6,000[3]	20,000[4]	20,000[4]	6,000[3]

1) Threshold in kilos = 0.
2) Allowed deviation in %.
3) Threshold value in kilos = 5.
4) Threshold value in kilos = 10.

2004:01 Hjälpverksamhet. Avrapportering av projektet Systematisk hantering av hjälpverksamhet

2004:02 Report from the Swedish Task Force on Time Series Analysis

2004:03 Minskad detaljeringsgrad i Sveriges officiella utrikeshandelsstatistik

2004:04 Finansiellt sparande i den svenska ekonomin. Utredning av skillnaderna i finansiellt sparande Nationalräkenskaper, NR – Finansräkenskaper, FiR Bakgrund – jämförelser – analys

2004:05 Designutredning för KPI: Effektiv allokering av urvalet för prismätningarna i butiker och tjänsteställen. Examensarbete inom Matematisk statistik utfört på Statistiska centralbyrån i Stockholm

2004:06 Tidsserieanalys av svenska BNP-revideringar 1980–1999

2004:07 Labor Quality and Productivity: Does Talent Make Capital Dance?

2004:08 Slutrapport från projektet Uppsnabbning av den ekonomiska korttidsstatistiken

2004:09 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

2005:01 An ignorance measure of macroeconomic variables

2005:02 Svenska hälsoräkenskaper. Ett system framtaget inom ramen för de svenska nationalräkenskaperna

2005:03 The sample project. An evaluation of pps sampling for the producer and import price index

2005:04 Finansiellt sparande i den svenska ekonomin. Utredning av skillnaderna i finansiellt sparande. Nationalräkenskaper, NR – Finansräkenskaper, FiR. Slutrapport

2005:05 Net lending in the Swedish economy. Analysis of differences in net lending National accounts, NA – Financial accounts, FA. Final report

2005:06 Varför får NR motstridiga uppgifter? Statistikens samanvändbarhet studerad inom ramen för Nationalräkenskaperna

2005:07 Evaluation of the possibility of producing statistics on production in the construction industry on a monthly basis