



Statistiska centralbyrån
Statistics Sweden

Foreign Trade and Industry Indicators Unit
Economic Statistics Department

Final report
2016-01-22

Evaluating and improving the validation
process in the Swedish foreign trade of
goods statistics

Grant agreement 06151.2014.004-2014.522

Evaluating and improving the validation process in the Swedish foreign trade of goods statistics



Summary

This report summarizes the activities conducted during the grant project 'Evaluating and improving the validation process in the Swedish foreign trade of goods statistics' at Statistics Sweden in 2015.

During the project, the validation and editing processes of important parts of the Swedish international trade in goods statistics (ITGS) were analysed and suggestions for improvements were made. For some parts improvements could be made within the project, whereas others require more work before the changes can be implemented.

Two areas in which improvements could be made during the project are the net trade balance and Extrastat. These changes will lead to more efficient workflows, better validation processes and better quality in the statistics. For the net trade balance the changes will lead to more focus on the most important providers of statistical information (PSI). For Extrastat a new step in the validation process has been implemented focusing on CN8-codes with large potential errors.

The suggestions not implemented during this project will be implemented in the grant project 'Improvements of tools and methods for ITGS' during 2016.

Table of content

Summary	2
1. Introduction	4
1.1 Background and objectives	4
1.2 Participants	4
1.3 Organisation and follow-up of work	4
1.4 Risk analysis	5
2. Activities and results	5
2.1 Use of VIES and VAT in the validation of Intrastat data	6
2.1.1 The current editing and validation process.....	6
2.1.2 Differences between Intrastat and VIES	8
2.1.3 How the sources relate to each other	9
2.1.4 Potential uses of VIES in the validation/editing process.....	10
2.2 Price validation of Intrastat	11
2.2.1 The current micro-editing of Intrastat.....	11
2.2.2 Suggestions for further improvements	11
2.3 Price validation of Extrastat	12
2.3.1 Current validation.....	12
2.3.2 The price validation method	12
2.3.3 Testing the price validation	15
2.3.4 Suggestions for further work	15
2.4 Macro-editing	16
2.4.1 Alternative macro-editing method.....	16
2.4.2 Macro-editing of Intrastat.....	17
2.4.3 Macro-editing of Extrastat	18
2.4.4 Suggestions for further work	19
2.5 Net trade balance	20
2.5.1 Intrastat editing	20
2.5.2 Extrastat editing	21
2.5.3 Suggestions for further work	21
2.6 Specific movements of goods	22
2.6.1 Current process	22
2.6.2 Collection, validation and checking	23
2.6.3 Future improvements.....	23
3. Conclusions.....	24

1. Introduction

There is a lot of ongoing work within the EU to modernise the statistical system for intra-EU trade in goods (Intrastat) with the aim to reduce the administrative burden on providers of statistical information (PSIs) and to increase the overall quality of the statistics. A large part of the work is undertaken in Eurostat's project 'Re-design of Intrastat' which is investigating different options for reaching the goal of burden reduction and quality improvements.

Within the context of the ongoing work many Member States are undertaking own projects to improve the quality of the international trade in goods statistics (ITGS). Improving the quality could become even more important in the future if Member States would compile their statistics using other Member States' data or use other Member States' data as auxiliary information.

1.1 Background and objectives

This project focuses on the validation and editing in the Swedish international trade of goods system. The overall objective is to investigate if and how the current methods and practices in the validation process can be improved in order to enhance the quality of the statistics.

The most important foreseen outcomes of the project are improved methods and practices in the validation process leading to a more efficient workflow and improved overall data quality.

Improving the quality of the statistics is important for the Swedish users, both internal and external, and may in the future also be important for other Member States.

1.2 Participants

The project was carried out by a project group with participants from three units at Statistics Sweden.

From the Foreign Trade and Industry Indicators Unit at the Economic Statistics Department: Jennie Bergman (project manager), Maria Adolfsson, Ari Mansikkaviita, Sofia Nilsson, Runo Samuelsson, Binniam Kidane and Peter Niemi.

From the Method Unit Enterprises and Organisations at the Process Department: Magnus Ohlsson, Can Tongur and Frank Weideskog.

From the Business Data Collection at the Data Collection from Enterprises and Organisations Department: Johanna Haller and Marie-Louise Janryd.

1.3 Organisation and follow-up of work

At the start of the project, the project group met to plan the work. The whole project group also met at three meetings during the year to follow up on the work.

For the work to be carried out as efficiently as possible the project group was divided into six smaller activity groups, one for each of the activities in the project. Each activity group had one activity leader with the main responsibility for the work. Staff from the methods unit participated in all activity groups, whereas staff from the foreign trade and data collection unit was involved in five and two activities respectively.

1.4 Risk analysis

A risk analysis was done after the first project meeting by one person from each unit within the project group and the project manager. The group found eight main risks. Five of the risks were connected to human resources; project member having to prioritize other tasks, difficulties to collaborate between the units due to different schedules, project members and/or project manager being absent for a longer period of time and not having enough time per activity. The other risks highlighted were not getting the right contacts with the Swedish Customs, the goals of the project not being concrete enough and lack of information of current methods and work processes.

Risk mitigation, responsible person(s) and when the risk mitigation should take place were discussed for each risk. For some of the risks the responsibility of risk mitigation was seen as a responsibility for the whole project group or activity group. For other risks, the project manager and the activity leader had the responsibility of risk mitigation.

For some risks the risk mitigation was seen as ongoing during the whole project, for example the risks connected to human resources, while for others the risk mitigation should be discussed at the first activity meeting or if the risk occurred.

2. Activities and results

The project was split into six main activities connected to the current validation and editing process of ITGS at Statistics Sweden.

The six activities were:

1. Map the potential use of VIES in the validation of Intrastat and evaluate, and if possible improve, the VAT validation of Intrastat
2. Study the possibility of removing time inefficiencies/possible bottle necks in the price validation of Intrastat
3. Study the possibility of using price validation for Extrastat as an auxiliary help in the aggregated validation process
4. Improve and evaluate a method already developed but not yet implemented for macro-editing (CN2-, CN4-, and country level)
5. Investigate if improvements could be made in the validation during the work with the net trade balance (rapid statistics)
6. Map and evaluate the current validation process of specific movements of goods to see if/how it could be improved

2.1 Use of VIES and VAT in the validation of Intrastat data

This activity examined the possibility of using data on VIES (VAT Information Exchange System) in the validation process of statistics on intra-EU trade.

Several projects have been carried out at Statistics Sweden to investigate the characteristics and potential of VIES data. This activity adds to these previous projects by focusing on the validation process and exploring potential uses of VIES data in this important step of the production process. To provide coherent analysis the report on this activity is structured as follows. An overview of the current validation process is first provided. Some ideas on how VIES data could be incorporated in the editing process are then highlighted.

2.1.1 The current editing and validation process

Data is continuously assigned to the Intrastat system and controls for identical posts are carried out. The validation process then proceeds in the following steps:

1. Control for non-response
2. Automatic imputation
3. Validation editing
4. Editing of individual PSIs
5. Price/micro-editing
6. Macro-editing
7. Editing using the VAT declarations
8. Editing for incomplete declarations
9. Labelling for publication and estimating for non-response
10. Output editing

The first step consists of a **control for non-response** and involves a follow-up of the non-response of the reference month and the five previous months. The penalty register is updated and if a report has not been received from a PSI a penalty is distributed.

The second and third steps of the editing process concern **automatic imputations and validation editing**. A validation error occurs if the PSI has reported an invalid value (for example a country or commodity code that is not part of the nomenclature). Validation errors are not present in the data reported electronically since the data is validated before the PSI can transmit the report. These reports can still contain incorrect information but not validation errors. Hence, validation errors are only present for the few PSIs which have been given permission to continue to submit paper declarations. For these reports validation errors concerning commodity items of small values are automatically imputed while validation errors concerning larger values are manually edited.

The data is then **edited at the level of individual PSIs**. In this step values reported by the PSIs are assessed individually. A validation list is compiled of suspicious values with high potential impacts on the data to

be published. Each value is attributed a score based on the impact and degree of suspicion. A threshold is then applied so that only critical values are edited.

Step five consists of the **price/micro-editing** process. In this validation process the unit prices are compared using values, weight and supplementary units. The potential errors are categorized according to different conditions and are then investigated. The score indicates the severity of the potential error. All newly reported observations are assigned a score by the validation program. See section 2.2. for more information about the price editing.

The sixth step concerns **macro-editing** the observations. A SAS program is run for the reference period. Excel files are generated that contain three spreadsheets (CN4, CN2 and country) for each flow and are subject to editing. Data for the past 13 months are included in each spreadsheet. A number of conditions are applied to select suspected critical errors and thresholds are applied to make sure that the errors that appear on the list are of a certain magnitude. Suspected errors are validated by contacting the enterprises in question. For more information regarding macro-editing see section 2.4.

The data is then edited by using the information submitted by the PSIs on the **VAT declarations**. The values in box 20 on the VAT declaration titled 'Intra-EU acquisitions' and box 35 titled 'Intra-EU sales' should correspond to the values reported in the Intrastat declarations for arrivals and dispatches. In practice, however, there may be some discrepancies between the two data sources due to aspects such as credit notes. A PSI may be contacted if there are large differences between the two sources. PSIs are only contacted if they are obliged to report information on Intrastat and their monthly trade exceeds SEK 10 million for one flow. However, Statistics Sweden may also contact enterprises that are not obliged to provide information if a value exceeding SEK 4.5 million of monthly trade is reported.

Editing of incomplete declarations is then carried out for PSIs that have only submitted parts of their total trade. Partial non-response such as missing commodity codes is included in this editing step. Systematic controls of the completeness of the submitted report are performed through validation on PSI level and by comparing the value reported in the VAT and Intrastat declaration. Correction of incomplete reports is carried out in the Intrastat system through manual amendments of total arrivals and dispatches for each PSI and reference period.

Step nine consists of transferring the data from the Intrastat system to a new database and **labelling** the data belonging to each publication with a specific notation (PubID). **Estimates for non-response** using an appliance in SAS are then carried out.

The final step entails **output editing**. The amendments for incomplete trade are checked and the values in the VAT declarations for PSIs that have not submitted a declaration (non-response) are validated. The correctness of the estimates on the level of the PSIs is checked for the largest companies in terms of value (PSIs with trade exceeding 30 million per month). In addition, estimates above 30 million SEK are controlled. Estimates on detailed level are then controlled by analysing the proportion of estimated trade for different combinations between commodities and countries. Finally, the largest estimates on CN8 level are checked.

2.1.2 Differences between Intrastat and VIES

Statistics Sweden receives information on VIES for both arrivals and dispatches. The data on dispatches is reported by Swedish enterprises while the data on arrivals is reported by enterprises in other EU countries exporting to Sweden.

VIES is reported to the Swedish Tax Agency on the 20th or 25th of the month following the month of reference using a periodic statement. VAT on the other hand is reported on the 12th two months after the reference month. It is only following the VAT declaration that the Swedish Tax Agency becomes aware of which companies are required to submit information on VIES using a periodic statement. The periodic statement is compiled manually by the enterprises. It is the responsibility of the individual enterprises to submit the periodic statement. Some enterprises are not aware of the obligation to submit information on VIES and therefore submit the periodic statement after the VAT declaration leading to some VIES reports being subject to late reporting.

The periodic statement is submitted by all enterprises that sell a good or service without VAT to an enterprise in another EU country. The statements are submitted on a monthly basis but companies that have a quarterly intra-EU sale totalling less than SEK 500 000 may report information on VAT on a quarterly basis. The report of sales of services without VAT to other EU-countries is carried out on a monthly or quarterly basis using the same form as the one used for trade in goods. The sources of the information on arrivals are the tax authorities in other Member States. The Swedish Tax Agency carries out a number of controls using the VAT statement on the information received from other Member States. According to the Swedish Tax Agency the quality of VIES information on dispatches is the same as the quality of the information on arrivals.

A control is carried out by the Swedish Tax Agency by comparing the VAT statement to the periodic statement. Differences between the two sources are often detected in this control. This difference may be due to currency effects and a sale or acquisition not being reported by the enterprise in question. The control is only carried out on sales and/or acquisitions of certain values. The thresholds that are used in determining

which observations that are included in the control procedure are confidential. The procedure can result in corrections and amendments of VIES data. The time needed for an enterprise to correct the data on VIES varies from case to case and according to the Swedish Tax Agency it is difficult to estimate the average time needed for corrections.

Another source of discrepancy between the two sources is the reporting enterprise not having a valid VAT number in the partner Member State. This is due to the fact that the buyer of the goods needs to be able to show that it has taxable person status in the other Member State but is not obliged to have a VAT number. It is therefore not possible to submit a periodic statement for all transactions of goods.

The Swedish Tax Agency believes that the information on VAT is more reliable than the information on VIES as there is more knowledge among enterprises on the reporting of VAT than VIES.

2.1.3 How the sources relate to each other

Here the correlation between the different data sources and an overview of how much the information in the VAT and VIES records can change from first point of data collection to later periods are presented.

Correlation between the data sources

Table 1 below shows that the correlation is very high between VAT and Intrastat data. For the time period 2013 to 2015 the correlation is 0.97 for both arrivals and dispatches, when estimating on PSI level. VIES and Intrastat data have a significant lower correlation with as low as 0.30 for VIES data on country and PSI level.

Table 1. Correlation between Intrastat, VAT and VIES, 2013-2015.

	Arrivals	Dispatches
VAT and Intrastat on PSI level	0.97	0.97
VIES and Intrastat on PSI level	0.85	0.51
VIES and Intrastat on country and PSI level	0.69	0.30

In table 2 below the correlation has been estimated for the same variables as above but with a simple random sample of 10,000 from the dataset used in table 1. The correlation for VAT and Intrastat data has similar values but the correlation between VIES and Intrastat data has changed. Looking at VIES data on country and PSI level the correlation is estimated at 0.96 for dispatches and 0.84 for arrivals. The reason for the large difference between the sample and the total is due to a few outliers in the VIES dataset that affect the correlation. This points to that VIES data needs to be edited with the help of VAT data before it can be used.

Table 2. Correlation between Intrastat, VAT and VIES for the simple random sample, 2013-2015.

	Arrivals	Dispatches
VAT and Intrastat on PSI level	0.97	0.98
VIES and Intrastat on PSI level	0.91	0.98
VIES and Intrastat on country and PSI level	0.84	0.96

Comparing the changeability in data sources

Statistics Sweden receives VIES data once a month which is stored in a database. The value is the latest, so in order to analyse how the data changes over time the VIES data needs to be stored away locally. During June to September 2015 VIES data was stored away so that values could be compared. The comparison for April 2015 is shown in table 3 below, which points to the conclusion that VIES data for arrivals has more changes from first data collection to later periods than VIES data for dispatches and VAT data for both flows.

Table 3. Percent of enterprises with changed values from previous period, reference month April 2015.

Reference month:	Dispatches		Arrivals	
	VAT, percent of enterprises with changed values	VIES, percent of enterprises with changed values	VAT, percent of enterprises with changed values	VIES, percent of enterprises with changed values
April				
Jun -> Jul	0.19	0.94	0.14	22.56
Jul -> Aug	0.14	0.64	0.12	4.14
Aug -> Sep	0.13	0.51	0.17	3.57

2.1.4 Potential uses of VIES in the validation/editing process

Statistics Sweden is currently building a new Intrastat IT-system. The new system will contain a box displaying information on VIES for each PSI. Currently, the VIES information is only available in SQL tables and not easily available in the IT-system. New information will therefore become available in the editing process when the new IT-system has been implemented.

According to the Swedish Tax Agency the quality of the information on VAT is of better quality than the information on VIES. To ensure a high-quality validation process, the data on VIES needs to be examined before being incorporated into the editing and validation process. An initial check needs to be carried out to check if the aggregate VIES data for a PSI corresponds to the total VAT data for that PSI. If the two sources show the same result, it is possible to incorporate the information on VIES into the validation process for the individual PSI.

When incorporating the information about VIES in the Intrastat IT-system, the new information will primarily concern the distribution of trade according to country on PSI level. It is possible to use this

information in the macro-editing process by controlling whether a PSI has traded with a certain Member State. This may serve to reduce the number of PSIs being contacted in the macro-editing process and so contribute to easing the overall burden on PSIs.

2.2 Price validation of Intrastat

This activity examined current price validation of Intrastat to see if the process could be improved and if time inefficiencies could be removed.

An overview of the current price validation of Intrastat process is first provided and then some suggestions for further improvements are highlighted.

2.2.1 The current micro-editing of Intrastat

Currently the micro-editing of Intrastat is done once a month and the starting point is a list of all the PSIs which have been flagged for a potential error. This list is then divided among the workers at the collection unit to be further investigated. The PSIs with potential errors that the editing staff find suspicious are sent a letter asking for clarification of the reported data.

The design of the letter is standardized so that every letter looks the same and has the same information, except for the data in question. For the purpose of this project we collected information over a three-month period in order to analyse how many of the PSIs reply to the letters.

Close to 600 letters were sent out during the time period and replies were received for 66 percent of the letters. Out of the received letters 43 percent led to a change. When the editing staff made an overview of the answers there were some values that looked suspicious enough to warrant a re-contact in which the reason for the contact was more thoroughly explained. When we got the responses from the second contact 89 percent of the replies led to a change.

2.2.2 Suggestions for further improvements

One conclusion is that we would save a lot of time if we were better at initially highlighting to the PSIs why we believe that the data may contain errors. This would lessen the risk that the PSI validates the data without really understanding why we are re-contacting them. The goal is to change the method of contacting the respondent from traditional mail to e-mail in this editing process. There are also plans to create an electronic identification method that does not require constantly new passwords which would make the login procedure simpler and more effective. In the web based collection tool (SIV) we will also make possible to add information in order to make it easier for the PSIs to know what to check.

This editing process is only possible to do once a month. The duration of time before being re-contacted is longer for the PSIs which report to Intrastat in the beginning of the collection period, compared to those

which report in the end of the collection period. The possibility of implementing a dynamic editing schedule in the new Intrastat IT-system has been investigated by the project members from the method unit. A few requirements will have to be met in order for this to happen:

- The editing staffs need to reorganize the system of sending letters to the PSIs. Currently, this is done once a month with the goal to send it out continuously by e-mail. A separate e-mail needs to be sent with password.
- Construct a system for how to divide the flagged values among the editing team. The ranking will fluctuate for the same value during the collecting period.

2.3 Price validation of Extrastat

This activity investigated the possibility of using a price validation in the editing process of Extrastat. The main objective behind testing this was to see if it could serve as an auxiliary help in the macro-editing at Statistics Sweden.

First a short introduction to the current validations and the idea behind a price validation of Extrastat is provided. Then the price validation method is presented. Finally, an overview of the testing of the method is given as well as suggestions for further work.

2.3.1 Current validation

In the current Extrastat validation process the data is validated at aggregated level with the help of an SAS program creating separate Excel files for imports and exports. These files include data on CN4-, CN2- and country level with flags indicating suspicious errors. The validation is good for finding errors on aggregated level but there is a risk of not finding errors on CN8-level.

Even though controls at microdata level are done by the Swedish Customs it seemed worth testing if price validation at Statistics Sweden could help to improve the quality of the statistics. The idea behind testing the price validation was to see if it could help in creating a more efficient work progress by pointing out the suspicious errors on microdata level.

2.3.2 The price validation method

Calculating suspicion

Since price validation is already done for Intrastat there was no need to develop a new method. The price validation only needed some adjustments based on the special characteristics of the Extrastat data.

In price validation, unit prices are compared in terms to value, weight and supplementary unit. In this editing process one criterion for selecting data to be verified is the potential impact on the summed values of trade in the published tables. Another criterion is the calculated suspicion that a value is erroneous.

The suspicion is defined as the distance between an observation and the closest of the upper and lower quartiles divided by the inter-quartiles distance. Since ratios like unit prices by nature have skewed distributions we use the logarithm of the unit prices.

$$\text{Suspicion}_i = \begin{cases} \frac{\log(\text{UP}_{Q_1}(i)) - \log(\text{UP}_i)}{\log(\text{UP}_{Q_3}(i)) - \log(\text{UP}_{Q_1}(i))} & \text{if } \text{UP}_i < \text{UP}_{Q_1}(i) \\ \frac{\log(\text{UP}_i) - \log(\text{UP}_{Q_3}(i))}{\log(\text{UP}_{Q_3}(i)) - \log(\text{UP}_{Q_1}(i))} & \text{if } \text{UP}_i > \text{UP}_{Q_3}(i) \end{cases}$$

For all else the suspicion is considered as equal to zero. When the quartile distance is zero the denominator is replaced by a fixed value or a value proportional to $\text{UP}_{Q_2}(i)$, the median unit price of similar goods.

Calculating the potential impact

In order to detect errors in the data that have significant potential impact on the results we start out from the difference in the statistical value (SEK) between observed value and an expected value, given the quantity. The median of unit prices is used, $\text{UP}_{Q_2}(i)$, multiplied by quantity_i as the best expected value. Notice that an error in the variable quantity_i results in a potential impact measured in value. An erroneous observation has potential impact on several domains of study in the published database. Therefore we first have to construct a potential impact variable for each domain. The Potential impact is the ratio of estimated error to the 'expected' sum for the domain of study.

$$\text{Potential impact}_i^g = \frac{|\text{Statistical value}_i - \text{Quantity}_i \times \text{UP}_{Q_2}(i)|}{\sum_{k \in g} \text{Statistical value}_k^*} \text{ if } i \in g$$

where g denotes a domain of study.

The median value $\text{UP}_{Q_2}(i)$ is computed on a homogenous set of data, independent of g , that makes the median a good predictor of the unit price for the object i . The sum $\sum_{k \in g} \text{Statistical value}_k^*$ is a sum over 24 months of statistical values for the domain of study g .

In order to determine the relative importance on the output five classification variables are being used: total arrivals/dispatches, 2-digit and 3-digit SITC, 6-digit CN and a set of important 8-digit CN-codes. The importance of these five factors is defined by a relative factor $O_v = 1$ to 5. Total arrivals/dispatches are given the factor $O_1 = 0.1$ and the other four higher values are set accordingly as of the relative importance of different levels of aggregations.

$$\begin{aligned}
 & \text{Potential impact}_i = \\
 & \text{maximum}_{\text{over } v=1-5} \left\{ \frac{| \text{Statistical value}_i - \text{Quantity}_i \times UP_{Q2}(i) |}{\sum_{k \in g_v} \text{Statistical value}_k^*} \times \frac{1}{O_v} \times f^{10 \log(\sum_{k \in g_v} \text{Statistical value}_k^*)} \right\} \\
 & = | \text{Statistical value}_i - \text{Quantity}_i \times UP_{Q2}(i) | \times \text{maximum}_{\text{over } v=1-5} \{ R_V(i) \} \\
 & \text{where } R_V(i) = \frac{1}{\sum_{k \in g_v} \text{Statistical value}_k^*} \times \frac{1}{O_v} \times f^{10 \log(\sum_{k \in g_v} \text{Statistical value}_k^*)}
 \end{aligned}$$

The one that has the maximum impact will be used when calculating the score-function.

Calculating the score

To calculate the score for a single object the following formula is used:

$$\text{Score}_i = \text{Suspicion}_i \times (\text{Potential impact}_i)^{P_{Imp}}$$

The object with the highest score gets the ranking number 1 and should be the first observation to investigate.

Figure 1 below illustrates that the boundary of the acceptance region is a line in the log-scale for Suspicion and Potential Impact. The slope is $-1/p_{Imp}$. The ordinate in origin depends on the number of observations we can afford to verify. The symbol 'x' represents a hit, i.e. an error in the data, and dots above the line represents outlier data that could not be verified as erroneous.

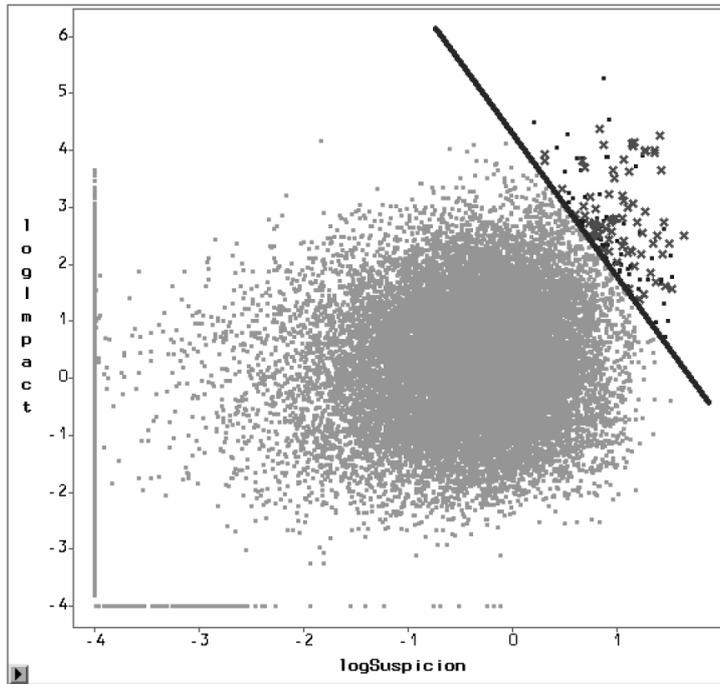


Figure 1. Acceptance region.

2.3.3 Testing the price validation

The price validation was tested by two employees at the unit for foreign trade in goods in September and October 2015. The test was done during the ordinary validation process, before the macro-editing to make the validation as efficient as possible.

The new SAS-programme for the price validation in Extrastat created one Excel file with around 1,500 lines. The variables used in the validation were:

- flow
- corporate registration number
- CN8-code
- country code
- value
- weight and/or supplementary unit
- kilo price/price per supplementary unit/weight per supplementary unit
- error type
- rank

Since the program created one Excel file for both exports and imports the list was divided into two in order to separate the flows. The lists were then ordered on lowest rank and then the work could begin. The top 20 lines for each flow were checked and some extra checks were done for lines with high values.

This new step in the validation process was not very time consuming and since the information was on microdata level the suspicious errors were easy to check. Further, some of the CN2 and CN4 codes flagged in the macro-editing were already checked in the price validation leading to a more efficient macro-editing.

The price validation was seen as useful and efficient. The potential errors included in the lists were relevant and with an impact on the statistics on detailed level.

2.3.4 Suggestions for further work

After having tested the price validation for two reference periods it was implemented as a new step in the validation process of Extrastat. We believe that it can improve the quality on the statistics without requiring a heaving workload.

However, some further developments are still needed to improve the method and the changes will be done during 2016. The developments include having separate programs for imports and exports and excluding procedure codes not included in the statistics.

2.4 Macro-editing

This activity included an evaluation and, if possible, an improvement of the current macro-editing of Intrastat and Extrastat. A new method for macro-editing of Intrastat was developed during a project in 2011 but never implemented. During this project, the new method was adjusted to be used also for macro-editing of Extrastat. The method was tested for both Intrastat and Extrastat by the staff at the unit for foreign trade in goods.

This section of the report is structured as following. The alternative macro-editing method is presented along with an overview of the current work and testing of the new method for Intrastat and Extrastat. The overview is done separately for Intrastat and Extrastat since these two data sources are edited separately. Finally, some conclusions and suggestions for further work are given.

2.4.1 Alternative macro-editing method

The concept with the alternative method is that you want to investigate numbers that lie outside the normality in combination with how much they affect the statistical output. In order to quantify that the following steps are taken:

1. Take the latest 24 periods and take sum and mean on CN2_i, CN4_i and country_i for weight, value and supplementary unit for each respective period.

$$2. \text{Suspicion}_i = \begin{cases} \frac{\log(\text{UP}_{Q1}(i)) - \log(\text{UP}_i)}{\log(\text{UP}_{Q3}(i)) - \log(\text{UP}_{Q1}(i))} & \text{if } \text{UP}_i < \text{UP}_{Q1}(i) \\ \frac{\log(\text{UP}_i) - \log(\text{UP}_{Q3}(i))}{\log(\text{UP}_{Q3}(i)) - \log(\text{UP}_{Q1}(i))} & \text{if } \text{UP}_i > \text{UP}_{Q3}(i) \end{cases}$$

3. To calculate the effect a value have in its own group the following formula is used:

$$\text{Effect}_i = \frac{|\text{value}_i - \text{value}_{Q2}(i)|}{\text{value}_{\text{mean}}(i)}$$

4. Calculate the score-value: Score = Suspicion_i × Effect_i

For CN2, CN4 and country the dataset is ordered respectively and ranked depending on the score-value. The highest ranks in the different groups should be investigated.

¹ Improving macro editing in Intrastat, grant agreement No. 20722.2010.001-2010.443.

2.4.2 Macro-editing of Intrastat

Macro-editing of Intrastat consists of checking data on an aggregate level. The data is checked for on chapter level (CN4, CN2) and country level. The checks are carried out for the reference month and the period that is subject to revisions (usually three months). Since there are far too many CN4 levels to check we select potential errors and merely check them. This is also done on CN2 level but not on country level (all country aggregates appear on the list but potential errors are highlighted).

The current process of macro-editing is carried out in the following steps: The method checks for suspected errors in aggregates at CN2, CN4 and country levels. These suspected errors are marked as 'high value', 'high weight', 'high supplementary unit' etcetera on a list in Excel. This gives us three lists per flow, one for CN2, CN4 and country respectively. For each aggregate value, weight and supplementary unit are shown for thirteen months. The editors then check all possible errors, contact PSIs with suspected errors, correct the errors that need correcting on microdata level and mark the suspected errors on the macro-editing lists. We mark 'corrected value' (2), 'not an actual error' (1) and 'uncertain' (0).

The new method of macro-editing was carried out in parallel with the ordinary macro-editing procedure in Intrastat for the reference months of July and August 2015. The new method was evaluated by a team that performs macro-editing on a regular basis. A number of tables comparing the two macro-editing procedures were compiled and analysed. Given that the two macro-editing procedures were only carried out in parallel two times no decisive conclusions could be drawn. However, the following observations were made:

- The ordinary macro-editing procedure currently used produces a list that can be analysed within the time constraints of the macro-editing process. The new macro-editing procedure is more time-consuming as the formatting is not as user-friendly as the ordinary procedure.
- The number of suspected errors that are corrected was modest for both reference months and both flows (less than 10 for all categories combined).
- The majority of corrected suspected errors were errors in supplementary unit.
- The majority of corrections were carried out on the CN4 level.
- It is more difficult to find the source of the suspected errors if the system flags a potential error of a unit being too low.
- The two macro-editing processes yield approximately the same results on CN2 level for both arrivals and dispatches. In fact, for the reference month of July 2015 the suspected errors that were ranked the highest for the category "weight" appeared on both lists. For the remaining months more than 7 of the 10 highest

ranked errors on the new lists appeared on the ordinary lists. The value-added of performing an additional macro-editing process on the CN2 level therefore appears low.

- Several errors that do not appear on the ordinary lists on CN4 level and country level do appear on the new lists for both arrivals and dispatches. On average around 5 of the highest ranked errors on the new lists appear on the ordinary lists for CN4. On country level the number of the 10 highest ranking errors on the new list that also occur on the ordinary list is even lower (around 2 for arrivals and 3 for dispatches). The new macro-editing method therefore appears to have potential in highlighting new errors on the CN4 and country level.

2.4.3 Macro-editing of Extrastat

The macro-editing of Extrastat is similar to the macro-editing for Intrastat in many ways. The data is checked at CN2, CN4 and country level as well as for the reference month and the revision period. However, for Extrastat all codes on CN2 level appear on the lists but not all countries. The twenty largest countries in terms of trade value are always included in the lists and the rest are dependent on potential errors.

The current process works in the same way as for Intrastat, with some exceptions. All suspected errors in the lists are checked and comments are written for each of the flagged lines. If a suspected error has an impact on aggregated level the editors can temporarily correct it. However, all potential errors are checked with the Swedish Customs which are responsible for contacting the enterprise and validate the suspected error.

The new method was carried out in parallel with the ordinary macro-editing in Extrastat for the reference months of August and September 2015. This differs from the reference periods for Intrastat due to data being available sooner from the Swedish Customs than from the PSI reporting data to Intrastat. Just as for Intrastat the new method was evaluated by staff performing macro-editing on a regular basis and the two different macro-editing methods were compared.

The following observations were made for the new macro-editing process for Extrastat data:

- In some cases, the new method gave high ranking to aggregates with low values and hence low impact on the total exports and imports.
- The new lists contained no information of the value, weight and supplementary unit. This is seen as helpful in the validation process, especially when the deadline is close and there is a need to prioritize what to check.
- Customs procedure codes which are not included in the statistics should to be excluded from the lists.

- If the method is to be used as a complement to the current method it might be enough with only the last month. If the method is to replace the current method also months subject to revisions need to be included.
- The new method was seen as quite time consuming.
- The experiences from the testing were a bit different depending on which flow that was examined. For exports, the list on country level was found to be of less importance and should not be prioritized whereas for imports the list on country level was considered good. The list on CN2 level was seen as the list of most importance for exports.

The new method also gave rise to some discussions regarding the focus and aim of the macro-editing.

2.4.4 Suggestions for further work

Based on the evidence presented above it is possible to suggest a new procedure for macro-editing of Intrastat. The suggested approach consists of the following steps. The five highest ranking codes on CN4 and country level for all categories (weight, value and supplementary unit) for both arrivals and dispatches are selected on the new validation lists. The ordinary validation lists are then controlled for these codes. If the codes do not appear on the ordinary validation lists they are added to the ordinary editing procedure and attributed historical values, weights and supplementary units as well as the appropriate indicator from the new lists. The ordinary macro-editing procedure is then carried out. This approach allows the ordinary editing procedure to be combined with highest ranking suspected errors of the new approach. This approach would benefit from a technical solution.

The approach described above could be extended by performing it for all months that are subject to revisions (usually three). This is, however, dependent on a technical solution as manually extracting several lists for each flow is resource consuming. If a technical solution cannot be implemented the new lists on CN4 and country level can serve as additional information in the editing procedure.

For Extrastat, the new method for macro-editing was found useful and seen as able to improve the data quality in the published statistics. Comparing the new method with the ordinary method for macro-editing showed that the new method found potential errors not included in the ordinary lists. The new method was found to be a good complement to the ordinary method, but not able to replace the method currently used.

Before implementing the new lists as a complement to the method currently used some changes have to be made. These changes include making the lists more user-friendly, excluding procedure codes not included in the statistics and adding time series in some way. One way to

incorporate time series for the most important errors could be to use the same approach as proposed for Intrastat. There is also a need for a discussion with the Swedish Customs regarding how to prioritise the potential errors in Extrastat.

2.5 Net trade balance

This section of the report focuses on the validation of the Swedish net trade balance (rapid statistics). The main aim was not to create new validation methods but to improve the current methods and make them more user-friendly.

First a short introduction to the Swedish net trade balance is given and then the editing of Intrastat and Extrastat is presented separately. Finally, some suggestions for further work are highlighted.

The net trade balance is estimated each month using rapid aggregated statistics, which are published one month earlier than the detailed statistics of the same reference month. As in the detailed statistics only Intrastat needs to be estimated, since total Extrastat figures are provided from the Swedish Customs.

2.5.1 Intrastat editing

Today, data editing of rapid statistics in Intrastat takes place in two rounds. The first round takes place in connection with the first time entry of data received into the system (transfer 1). Most of the editing focuses on PSIs with large deviations from earlier reported values, as well as checking of the 20 most prioritized PSIs, in addition to those PSIs which have changed their organisation number. The second round focuses on PSIs with large deviations from earlier reported values and takes place in connection with entry of newly received data (transfer 2).

In transfer 1, the total arrival and dispatch value for each PSI for the current publication month is being edited through five different data checking procedures (check 1-5 below).

- Check 1: PSIs with deviations > SEK 50 million from the average value of the four preceding months.
- Check 2: PSIs with deviations < - SEK 50 million from the average value of the four preceding months.
- Check 3: PSIs with high values, covering both large absolute deviations and percentage deviations from the average value of the four preceding months, and which are not identified by check 1.
- Check 4: PSIs with low values, covering both large absolute deviations and percentage deviations from the average value of the four preceding months and which are not identified by check 2.
- Check 5: Concerns PSIs which have had some form of organisational change over the last four months. If reports are

received for a newly registered PSI which does not have any data in the detailed statistics (t-2), this value is manually transferred to the former PSI. Possible corrections during the checking are made in the rapid statistics system.

About one week later transfer 2 is carried out for the current month. After this transfer is completed, check 1-4 are carried out for the new PSIs, as well as check 6 - 7:

- Check 6: Concerns the largest and most important Intrastat PSIs (priority 1-20 for arrivals and priority 1-20 for dispatches).
- Check 7: Final checking for each PSI (checks 1-4 as above).

Table 4 below shows the number of PSIs in check 1-5 for January to June 2015. Check 6 and 7 are excluded from the table since check 6 always contains 40 PSIs and check 7 only contains a few PSIs.

Table 4. Number of PSIs in data check 1-5, January to June 2015.

Period	Arrivals, check					Dispatches, check				
	1	2	3	4	5	1	2	3	4	5
201501	16	37	35	76	10	12	31	13	41	10
201502	23	30	49	50	10	12	19	24	38	8
201503	28	24	79	50	12	35	17	53	18	8
201504	18	17	46	54	15	15	17	32	31	8
201505	18	22	36	59	16	11	23	36	31	10
201506	28	17	54	50	15	26	10	32	21	10
Average	22	25	50	56	13	18	19	32	30	9

When the editing processes have been completed, a programme in SAS is run which estimates the total value for the current month.

2.5.2 Extrastat editing

Records from the Swedish Customs are entered into the Extrastat database, and the data editing starts. In parallel with the macro-editing, the 20 largest commodity items (CN8-level) in terms of value for imports and exports of the current publication month are checked. If the commodity items are suspected of being incorrect, the Swedish Customs are contacted and necessary changes are made in the Extrastat database. See section 2.4.2 for more information about macro-editing of Extrastat.

2.5.3 Suggestions for further work

The aim of this activity was partly to rationalize the data editing process used for rapid statistics in Intrastat. A relatively simple analysis was carried out into data check 3 and 4 (described above) concerning the number of PSIs checked over a 6 month period (January - June 2015). Arrivals and dispatches were considered separately.

As displayed in table 5 below, it appears as if the number of PSIs in data check 3 and 4 during a regular production month amount to 106 on

average for arrivals and 62 for dispatches. In an attempt to reduce the number of items, a lower limit value was set to SEK 15 million. The lower limit value excludes PSIs (that would otherwise be flagged for investigation because the reported value is highly suspicious) with a reported value that deviates less than SEK 15 million from the average value of the four preceding months, and which are not identified by checks 1 and 2. On average, the test shows that the limit would reduce the number of PSIs which were flagged during the 6 month period by more than 40 percent without considerable impact on the total value of trade. Arrivals would tend to be overestimated by SEK 26.5 million, whilst dispatches would be overestimated by SEK 4.6 million. However, the total value of imports amounted to SEK 95.8 billion, while exports were valued at SEK 99.0 billion.

Table 5. Number of PSIs in check 3 and 4, January to June 2015.

Period	Current			New		
	Arrivals	Dispatches	Total	Arrivals	Dispatches	Total
201501	111	54	165	59	29	88
201502	99	62	161	59	39	98
201503	129	71	200	80	41	121
201504	100	63	163	60	37	97
201505	95	67	162	49	33	82
201506	104	53	157	65	29	94
Average	106	62	168	62	35	97

Furthermore, manual handling tasks could be reduced by using less Excel worksheet tabs when carrying out data check 1-4 and 7, and by exporting SAS output into Excel instead of printing on paper.

A SAS solution has been developed to capture and avoid repeated values in the final editing for each PSI which were already flagged for investigation in transfer 1 or 2.

2.6 Specific movements of goods

This activity focused on the editing process of the Swedish specific movements of goods. The aim was to map and evaluate the current process to see if/how it could be improved.

This section of the report gives a brief introduction to specific movements of goods with a special focus on the validation and editing. Some ideas for future improvements are also given.

2.6.1 Current process

Specific movements of goods consist of vessels and aircraft, deliveries to vessels and aircrafts, electricity and industrial plants. Currently, the data collection of specific movements of goods is processed in a separate Access database. All reporting is done on paper forms.

As data collection for specific movements of goods differs from the ordinary collection, it is worth mentioning the data sources. These are specified in table 6 below.

Table 6. Sources for specific movements of goods.

Movement	Source
Vessels and aircraft, incl. space craft	Swedish Transport Agency (Maritime and Aviation Department)
Deliveries to vessels and aircraft	Statistics Sweden
Electricity	Swedish Energy Agency
Industrial plants	Statistics Sweden
Goods delivered to and from offshore installations	Statistics Sweden
Sea products	Swedish Agency for Marine and Water Management

2.6.2 Collection, validation and checking

Monthly incoming forms are manually registered and the editing is done during this process. Thereafter the data is exported into the Intrastat and Extrastat systems. After this, price validation and checking is processed in a similar way as described under 2.3 and 2.5.1., as well as control of time series, comparison of CN8, price unit, weight, value and supplementary unit.

Duplicate control for vessels and aircraft, as these are considered as unique objects, is done by comparing the Vessel Call Sign and Aircraft Registration Mark to corresponding, or otherwise similar, value/CN6-8/corporation registration number already reported to Intrastat or Extrastat.

Furthermore, in connection to the macro-editing of Extrastat, a special checking based on high values is done in co-operation with the Swedish Customs.

2.6.3 Future improvements

The new Intrastat IT-system will include specific movements of goods. If possible, also Vessel Call Sign and Aircraft Registration Mark will be included as extra variables helping to control duplicate values. Both the Vessel Call Sign and the Aircraft Registration Mark are today obtained through different sources making the editing process cumbersome.

In connection with the price validation of Intrastat, a list is produced once a month where reported value is flagged by editing rules. This list on suspected errors is also going to be included in the new IT-system where all values are gathered and are investigated with the help of a SAS-program.

The reporting, though, remains for the time being paper based with the goal to shift to web-based data collection. Investigating the shift is a part of the grant project 'Improvements of tools and methods for ITGS' for

2016. Hopefully, the Intrastat Data Entry Package (IDEP) can be used for PSIs reporting on a regular basis. By collecting data electronically, less manual work needs to be done. Inbuilt functions in an electronic reporting system alerts the PSI's at an instant if any value looks suspicious and could so contribute to reducing both editing time and cost.

Statistics Sweden receives data on sea products from the Swedish Agency for Marine and Water Management on a regular basis. The data is currently not included in the statistics. During 2016, a project will be carried out to evaluate the data and see if it can be used by ITGS and/or National Account.

3. Conclusions

Statistics Sweden receives both VAT and VIES data from the Swedish Tax Agency but currently only VAT is used in the validation process. The development of a new IT-system for Intrastat is ongoing and in the new system VIES data for each PSI will be displayed. This will make it possible to incorporate VIES in the validation process in a user friendly and easy way. VIES data could also be used in the macro-editing process to validate if a PSI has traded with a certain Member State. This could reduce the re-contacts with the PSIs and hence ease the overall burden.

The price validation of Intrastat could be improved by implementing a dynamic editing schedule and by improving the letters sent to PSIs to help them to understand why we are re-contacting them. Before the dynamic editing schedule can be implemented the new Intrastat IT-system has to be in place as well as new routines.

Price validation for Extrastat was tested with positive results as it points out potential errors not found in the macro-editing. The validation process for Extrastat has already been extended with this new step, however some adjustments in the method is still needed and will be taken care of during 2016.

The testing of an alternative macro-editing method gave somewhat divergent results. One problem was especially highlighted; the lack of time series. This problem could be dealt with by including the aggregates with the lowest ranks on the new list to the ordinary macro-editing list. However, this requires a technical solution. In lack of a technical solution, the new method could be used as a complement to the ordinary macro-editing method.

The editing of the Swedish net trade balance is done in several steps. In an attempt to reduce the number of PSIs being checked a lower limit was set in the validation. This reduced the number of flagged PSIs with more than 40 percent without having a considerable impact on the statistics. The change will lead to more focus being put on the values that have an impact on the statistics.

For specific movements of goods the goal is to change from paper forms to electronic reporting, hopefully IDEP, which will enable validation controls in the collection. With the help of the new Intrastat IT-system and new SAS-programs it will be possible to reduce the manual work and focus more on the validation.

The overall goal of this project was to analyse the current validation and editing process of the Swedish ITGS to see where improvements could be made. Improving the validation and editing could in turn improve the quality of the disseminated statistics. The project found areas within ITGS which could benefit from new methods or by improving the current.

There are two areas within ITGS in which improvements were made during the project; net trade balance and Extrastat. These changes will lead to more efficient workflows, better validation processes and better quality in the published statistics.

The suggestions for further improvements from this project will feed into the grant project 'Improvements of tools and methods for ITGS' for 2016 in which they will be implemented.