Improving macro editing in Intrastat

Grant Agreement No. 20722.2010.001-2010.443

Contact: Olle Håkanson, Statistics Sweden

E-mail: olle.hakanson@scb.se

Summary

This project's task is mainly twofold. One task is to evaluate the methodology of the macro editing in Intrastat and, if necessary, propose a new one. The other main task is to propose an idea of how to make an IT-system that will allow us to work with current data instead of a "frozen" list of suspected errors. The main objective is to improve the quality of the Swedish foreign trade of goods statistics by a more effective IT system and a methodology with higher ratio of corrected values compared to flagged errors.

Macro editing in Intrastat in Statistics Sweden is done on three aggregates, CN4, CN2 and country. This, obviously, gives us too many groups to check so we need a methodology that gives us a list of potential errors. The macro editing is the last main check before the estimations for non-response and the only check of its kind.

This project finds that the current methodology is somewhat ineffective in finding suspected errors and not as flexible as we would like. We therefore propose a new methodology based on the standard tool SELEKT. This methodology combines two aspects, suspicion and effect. Combining these two gives us a score that can be ranked.

By separating the data concerning macro-editing into two databases we can solve the problem with "frozen" data. One database will contain all the aggregates, their values and the score of suspicion and effect. The other database will contain the codes for how we treat the suspected errors and comments. The first database collects data from the main Intrastat micro database and can be reloaded as many times as we want for each month. In that way we can produce the list of potential errors whenever we wish to and therefore we have the possibility to work with all the data available.

Background

Macro editing is the last main check before the estimations for non-response and publishing work begins. The purpose of macro editing is to detect possible errors at aggregated levels and correct them on micro level. It is the only checking done on aggregated levels and will detect other kinds of errors than those done on micro level. It is therefore of vast importance that this check is of high quality.

The aggregated levels we check are CN4, CN2 and country. 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 (we check all country aggregates). The method to determine which deviations are considered to be a potential error is old and needs to be evaluated.

The current IT-system for macro editing is fairly new but somewhat unreliable in some aspects and not as flexible as we would want it to be. It is not written by professional IT-personnel but by the staff at the Foreign Trade unit, the personnel that do the actual editing. Therefore the current IT-system does not follow standards and would be difficult to support by other people then those who built it.

One of the main drawbacks with the current system is that it uses "frozen" data. At a given point in time we "freeze" the dataset and the macro editing begins. Late arrivals of intrastat reports are therefore not included in the macro editing. Also some micro editing is performed at the same time as the macro editing. If this micro editing causes changes in the Intrastat data it will not affect the "frozen" data used in macro editing. This problem with "frozen" data is not something unique for the macro editing in Intrastat, if we can find a suitable solution the ideas of that solution could be used in other checks.

Human resources used

The project began in February 2011 and was completed in November the same year. The work was carried out by Olle Håkanson and Runo Samuelsson from the Economic Statistics Department (Foreign Trade and Industrial Indicators Unit) and Can Tongur from the Process Department (Method Unit for Enterprise, Organisation, Real Estate and Environment Statistics) and Bengt Risberg from the Process Department (IT-Unit 3), all at Statistics Sweden.

Work process

The following main tasks were planned in the project:

- Writing of report.
- Evaluate the current methodology of the macro-editing.
- If possible propose a new methodology that gives higher ratio of corrected values compared to possible errors.
- Propose a new system that provides the possibility to work on current data instead of "frozen" data.
- If possible integrate this into the current IT-system.
- International benchmarking; investigate other member states macro-editing tools and methodology.

Due to personnel issues (one member of the staff could not participate in this project) the last task, International benchmarking, has not been done. We prioritized the other tasks.

The task of integrating the proposed solutions into the current IT-system will have to wait since we in the near future will construct a new IT-system for Intrastat. This report will however present what we plan to implement.

The current process of macro-editing works, briefly, as follows: The method checks for suspected errors in aggregates at CN2, CN4 and country levels. These suspected errors are marked as "High value", "Low weight", "High supplementary quantity" etcetera on a list. This gives us three lists per flow, one for CN2, CN4 and country respectively and values are shown for thirteen months. The editors then checks all possible errors, correct those who needs correcting on micro level and mark the suspected errors on the macro editing list. We mark "Corrected value" (2), "Not an actual error" (1) and "Uncertain" (0). The ratio of "Corrected values" over total suspected errors is one way of evaluating the current process and methodology.

The table below shows the total number of suspected errors (T) for each flow and CN-groups for 23 lists (April 2009 – February 2011). It also shows how these errors have been treated, if they were indeed actual errors or not or if we deemed them uncertain. We use the code for uncertain (0) for example if the data provider has not been able to inform us if the reported figures are correct or not in time. In each list we usually check for errors in four months, but in some lists we check 12 months (yearly revisions).

Flow	Group	T	0	1	2	Ratio T/2 (%)
Arrivals	CN4	1 936	72	1 600	264	13,6
Dispatches	CN4	1 897	153	1 588	156	8,2
Arrivals	CN2	2 590	23	2 427	140	5,4
Dispatches	CN2	2 463	69	2 281	113	4,6

The table shows us that CN4 has, as suspected, a higher ratio than CN2 but both are rather low. The absolute majority of flagged errors are indeed not actual errors.

To be subject for editing in the current method, the 13 month total (including the reference month) must be above a specific, fixed, value. For high values, the requirement is that the unedited value (of any of the three variables) exceeds 5 times the average value of 12 months for that specific domain. For low values, the control is if the unedited value is less than one tenth or one thirtieth of the 12 month total, depending on the level of the unedited value. Our experience is that the current method is perhaps too stiff and as shown above has a non-satisfactory hit rate.

Implementation

The proposed method

Statistics Sweden has for several years carried on concentrated work on improvement of selective data editing and has developed a generic tool, SELEKT, for micro data editing. The tool is now used in several economic surveys and in other surveys, the ideas behind the tool are applied instead of the complete tool itself. The main thought behind such a generic tool is to achieve some optimal level of editing by minimizing remaining bias with respect to different domains of study.

The selective editing method behind SELEKT can be considered as combining two leading components: suspicion and impact. The basic approach is to construct score functions for each variable of interest, may it be ratio variables, regular straight-expansion variables or derived variables corresponding to either one of them. The score function can be interpreted as yielding an expected impact for each variable by combining both suspicion and potential impact with respect to all possible domains of study. Suspicion is defined as the degree of deviation from an expected value (some mid-point of the distribution) or, more common, some upper and lower limits. Impact is defined as the influence on output, given that a specific reported item is flagged as erroneous.

The idea behind SELEKT spawns from micro-editing in Intrastat. For this macro editing project, we have applied the same micro data editing method but adapted to macro level data. The method is defined, in our case, as the following.

Let the unedited value be $y_{j,d}^{une}$, where *une* stands for unedited value and *j* represents any of the elements within domain *d*, i.e. a specific 4-digit code within the CN4 domain. The editing is of course divided by flows, into arrivals and dispatches. Let the corresponding expected value be $\tilde{y}_{j,d}^{\exp}$. From the distribution of *j,d* we compute the expected value which is the median, the lower quartile $Q1_{j,d}^{y}$ and the upper quartile $Q4_{j,d}^{y}$. We define *Suspicion* as:

$$Suspicion_{j,d}^{y} = \frac{Q1_{j,d}^{y} - y_{j,d}^{une}}{Q3_{j,d}^{y} - Q1_{j,d}^{y}} \text{ if } y_{j,d}^{une} < Q1_{j,d}^{y}$$

$$Suspicion_{j,d}^{y} = \frac{y_{j,d}^{une} - -Q4_{j,d}^{y}}{Q3_{j,d}^{y} - Q1_{j,d}^{y}} \text{ if } y_{j,d}^{une} > Q4_{j,d}^{y}$$

Whenever the unedited value is within the quartiles, suspicion is per definition zero. Potential impact is defined as the normalized deviation of the unedited value from the expected value:

$$Potential\ Impact_{j,d}^{y} = \frac{| y_{j,d}^{une} - \tilde{y}_{j,d} |}{(G(y)_{j,d})^{Oboe}}$$

The normalizing factor $G(y)_{j,d}$ is some function of data *j,d*, e.g. the mean for some time period or the sum. The exponent *Oboe* is a number above zero and maximally one (=1) and regulates the relative importance of larger commodity groups. We the compute a score function, interpreted as the anticipated impact, according to the following:

$$Score_{j,d}^{y} = Suspicion_{j,d}^{y} \times (Potential\ Impact_{j,d}^{y})^{Pimp}$$

As can be seen, the score function is a combination of the suspicion and the impact with an adjustment component to the impact (*Pimp*).

One should have in mind that macro editing differs greatly from micro editing in terms of performance, i.e. expected hit rate. In micro editing, a specific observation is

analyzed. For Intrastat, the micro editing at Statistics Sweden concerns the unit prices, i.e. the ratio between invoiced value and weight or invoiced value and supplementary quantity and not the value, weight or supplementary quantity themselves. Such an editing is of course more successful since the variation of unit prices is normally restricted by nature for most commodities (CN8). This desirable feature is of course not present in the case of macro editing in which we concentrate on invoiced value, weight and supplementary quantity separately. Since this editing is done by CN4, CN2 and Country, each divided on flow, we have 6 tabulations and in which the CN4 level is the most exhaustive. Having edited the CN4 level thus implies, more or less, that CN2 is already taken care of.

The proposed IT-solution

The main drawback of the current IT-system is, as stated before, that we work with "frozen" data. SAS creates an Excel-file that contains the possible errors at that point in time but we work with that file, or list, for a couple of days. During those days new reports are made and old are corrected both in micro and macro editing. There is a risk that we will miss some macro errors and that we will waste time checking already corrected errors. The main task of this project, in IT terms, is to provide a solution that enables us to work with current data all the time. This new solution should also have all the functionalities of the old one and fulfil some extra demands.

We plan to use two databases to solve the problem with "frozen" data. One database, let us call it the "Methodological", will contain all aggregates' values (invoiced value, supplementary quantity and weight for each flow) for any number of months and their score and also a ranking of the scores. We will also store what kind of error we suspect, i.e. high quantity, low weight etcetera. This database collects data directly from the main intrastat database where micro data is stored and is constantly updated. The other database, let us call it the "Editing", will store all the markings (0, 1, and 2) and comments for each aggregate subject for editing. This is the database that can be edited from the interface.

The interface will present all aggregates subject for editing and, as now, their values for the last 13 months. It will point out what values are suspected errors, what kind of error we suspect and the ranking of the error. The editor can then, as before, mark the aggregate with a 0, 1 or 2. The next time we update the list all 1:s should be excluded from the interface (if the values have not changed more than 5 %). This is a feature that exists in the current IT-solution and is very useful. In the end the editor will have a empty list or a list only containing 0:s (uncertain errors).

The "Methodological" database is the one collecting micro data from the main Intrastat data base. Since the "Methodological" and the presentational "Editing" databases are separate we can update the "Methodological" without it interfering with the list. We will not work with current and changing data, since that would be impractical, but rather choose when to update our lists. These updates can be done whenever we choose to and how often we choose to, every minute or once per day.

Other demands on the IT-system are that we should be able to set a number of parameter values at any time. These are:

- The number of months subject for editing

- Six ranking parameters, one for each of value, weight and quantity for both CN2 and CN4. All rankings below this will be presented in the interface and therefore subject to editing
- A fixed value for CN4, all invoiced values above this should always be shown regardless of its rank
- A fixed value for CN2, all invoiced values above this should always be shown regardless of its rank

All these parameter values will be stored either in a separate database for just the macro-editing or in a database with the other parameter values for Intrastat.

Conclusion

This project has resulted in an evaluation of the current method and a proposal for a new one. It has also produced an implementable idea of how we can work with current data instead of "frozen" lists.

A good aspect of the proposed IT-solution is that the methodological part (calculating the possible errors) and the presentational part (that the editors work with) are separate. We can revise the method, or change the software that calculates it, without changing the presentational part.

Continued work

The next step is to further evaluate the proposed method. A "shadow program" that produces a list for one month at a time is ready and tested. We will compare the current and proposed methods by checking what rankings the errors we corrected in the current list had. We will also examine if there are low ranking (highly suspected) groups in the new list that was not present in the current list. If all evaluations are satisfactory then we will implement as suggested in this report.

The results from this project can also be directly applied to the macro editing done on Extrastat figures. We will however not implement them both at the same time but rather work with the new process in Intrastat for a while and then evaluate it.