Statistiska centralbyrån Statistics Sweden

# A Selective Editing Method considering both Suspicion and Potential Impact, developed and applied to the Swedish Foreign Trade Statistics

2006:3

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

| | |
|---|---|
| 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 statistic |

Continued on inside of the back cover!

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

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

**Background Facts**

# A Selective Editing Method considering both Suspicion and Potential Impact, developed and applied to the Swedish Foreign Trade Statistics

**Economic Statistics 2006:3**

**Statistics Sweden**
**2006**

**Background Facts**

**Economic Statistics 2006:3**

# A Selective Editing Method considering both Suspicion and Potential Impact, developed and applied to the Swedish Foreign Trade Statistics

Statistics Sweden
2006

| | |
|---|---|
| Previous publication | Published since 2001 |
| Producer | Statistics Sweden, Macroeconomics and Prices Department<br>Box 24 300<br>104 51 Stockholm |

| | |
|---|---|
| Inquiries | Anders Jäder, tfn: +46 8 506 947 28<br>anders.jader@scb.se<br><br>Anders Norberg, tfn: +46 8 506 946 79<br>anders.norberg@scb.se |

When quoting material from this publication, please state the source as follows:
Source: Statistics Sweden, Background Facts, Economic Statistics 2006:3, A Selective Editing Method considering both Suspicion and Potential Impact, developed and applied to the Swedish Foreign Trade Statistics



ISSN 1650-9447

ISBN 91-618-1321-4
ISBN 978-91-618-1321-6

URN:NBN:SE:SCB-2006-X100ST0603_pdf

Printed in Sweden
SCB-tryck, Örebro 2006:3

# Preface

This report by Anders Jäder and Anders Norberg describes a new editing system which has been developed and implemented to the Swedish foreign trade statistics. The report was presented at the conference "Work Session on Statistical Data Editing" in Ottawa, Canada, May 16–18, 2005, which was arranged by United Nations Statistical Commission and United Nations Economic Commission for Europe[1].

The long series of conferences of this kind originate from an international cooperation in the field of editing that started in 1980. Statistics Sweden has participated from the start and our participants has found the conferences informative and enjoyable, and most relevant for Statistics Sweden.

Statistiscs Sweden, March 2006

Lars Melin

Anita Ullberg

[1] Minor changes have been made in the report after the conference. All documents presented at the conference can be found at http://www.unece.org/stats/documents/2005.05.sde.htm

# Contents

# 1 Summary

A score function computed as a weighted geometric mean of measures of suspicion and potential impact has successfully been implemented in the editing process of the Swedish foreign trade statistics. There are well over 10 000 statistical table sums to be produced and published each month. We have developed a formula with which the tolerable impact of the errors on the statistics on all aggregation levels and sizes of table sums can be expressed in one single variable. The survey managers have set the values of six constants that reflect the importance of potential errors on different aggregation levels and sizes of the sums.

The method needs relevant and accurate medians and quartiles for homogenous groups. For example, we have to decide whether we shall use current or historical data, the minimum number of observations to be used, and whether to use weighted or unweighted quartiles. In total we have ten different specifications to decide on to get the best possible performance. Hundreds of thousand different combinations of specifications were tested on raw and edited historical data

The sum of changes in invoiced value was 494 MSEK for the old method and 819 MSEK for the new method, when they were used in parallel in December 2003. The hit rate has increased from about 40 percent to 65 percent. Guided by this initial analysis the method was implemented in production in January 2004. Process data are now produced every month in a continuous search for best specifications.

# 2 Introduction

Intrastat is a survey created for Member States of the European Union (EU) that covers trade of goods among States within the EU. Each month some 350,000 data are collected from enterprises in a total census with cut-off, by Statistics Sweden. One editing process is the checking of unit prices. 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 and the other criterion is our *suspicion that a data value is erroneous*.

The foreign trade statistics are published for in- and outflow at the 6-digit classification according to the CN-nomenclature[2] and higher aggregation levels. Statistics are also published with the alternative item classification SITC[3] and for each of the countries in the EU. There are well over 10 000 table sums per month in the official database. In the paper we describe a formula for reducing the problem of deciding the tolerable impact on all these published statistics into a choice of six constants.

The method needs relevant and accurate medians and quartiles for homogenous groups. Some issues are: Should we use historical data or only current data, what is the minimum number of observations needed, should we calculate weighted or unweighted quartiles, is it necessary to split data into in- and outflow of goods? Detailed grouping is in conflict with demand on a minimum amount of data for computing quartiles and medians, a problem that also involves the number of months of historical data. In cases when there are enough data for detailed grouping, one issue is if one shall use information on all such levels or only on he most detailed level.

Statistics Sweden has saved raw and edited data since the year 2000. In the data there is information on 8-digit CN (i.e. more detailed than what is published), country, enterprise, year and month. We have tested the editing method with ten specification parameters on the data to find out which set of values is the best. Guided by this initial analysis the method was implemented in production in January 2004. We continuously run tests by changing some of the specifications almost every month. Process data are produced in a search for best specifications.

[2] The Combined Nomenclature has 8 digits. The first 2, 4 and 6 digits form relevant groups of goods.

[3] Standard International Trade Classification

# 3 Suspicion

The median and quartiles are fundaments in modern editing procedures, see Hidiroglou and Berthelot (1986). We define suspicion 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 take the logarithm of the unit prices.

Let $UP_i$ be the unit price for an observation *i* in our current data, i.e. $UP_i = \frac{Invoiced\ value_i}{Quantity_i}$.

Let $UP_{Q1}(i)$ and $UP_{Q3}(i)$ be the lower and the upper quartiles for unit prices computed on historical data – or computed on all the data for the current month – that belong to a "homogenous group", to which the observation *i* also belongs. The "homogenous group" is explained in sections 20-27.

We define *Suspicion* as:

$$Suspicion_i = \begin{cases} \dfrac{log(UP_{Q1}(i)) - log(UP_i)}{log(UP_{Q3}(i)) - log(UP_{Q1}(i))} & \text{if } UP_i < UP_{Q1}(i) \\ \dfrac{log(UP_i) - log(UP_{Q3}(i))}{log(UP_{Q3}(i)) - log(UP_{Q1}(i))} & \text{if } UP_i > UP_{Q3}(i) \end{cases} \qquad (1)$$

*Suspicion* is zero otherwise. When the quartile distance is zero, the denominator is replaced by a fixed value or a value proportional to $UP_{Q2}(i)$, the median unit price of similar goods.

# 4 Potential impact in a multi-purpose survey

In order to detect errors in data that have significant potential impacts on the results we start out from the difference in *Invoiced value* (SEK) between observed value and an expected value, given the quantity. We use the median of unit prices, *UPQ2(i)*, multiplied by *Quantityi* as the best expected value. Notice that an error in the variable *Quantityi* 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.

$$Potential\ Impact_i^g = \frac{\left| Invoiced\ value_i - Quantity_i \cdot UP_{Q2}(i) \right|}{\sum_{k \in g} Invoiced\ value_k^*} \quad \text{if } i \in g$$

where $g$ denotes a domain of study.

The median value $UP_{Q2}(i)$ 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 $\sum_{k \in g} Invoiced\ value_k^*$ is a sum over 24 months of *Invoiced values* for the domain of study $g$. When $g$ is a domain with varying total from month to month, without a stable seasonal pattern, we think it is better to use an annual total. If, on the other hand, the domain of study $g$ has a distinct seasonal pattern it might be better to compute this sum only on the current month and/or the same month the last years. To simplify the method we have made the choice to use annual data.

There are two reasonable demands on a comparison of acceptable impacts on two domains of study *g1* and *g2*:

- Size: If $g_1$ and $g_2$ are two domains of study formed by the same classification variables, for example by in-/out-flow and 2 digit SITC, we tolerate a relatively smaller impact of errors on $g_1$ if $g_1$ over the last two years has had a larger sum of trade than $g_2$.
- Importance of classification variable: If $g_1$ and $g_2$ are two domains of study formed by different classification variables, but $g_1$ and $g_2$ over the last two years have had the same size of the sum of trade, we tolerate a relatively smaller impact of errors when the classification variables are more aggregated. 2-digit SITC is more aggregated than 3-digit SITC and 6-digit CN.

After consultation with survey managers we have settled the importance of size by the constant *f* and the relative importance for the five classification variables 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 are defined by a relative factor *Ov*, *v*=1–5. Total arrivals/dispatches is given the factor *O1*=0.1 and the other five higher values are set subjectively by the management according to their view of the relative importance of different levels of aggregations.

We have constructed this measure of potential impact.

$$Potential\ impact_i = \underset{over\ v=1-5}{\text{maximum}} \left\{ \frac{\left| Invoiced\ value_i - Quantity_i \cdot UP_{Q2}(i) \right|}{\sum_{k \in g_v} Invoiced\ value_k^*} \cdot \frac{1}{O_v} \cdot f^{\,{}^{10}\log\left(\sum_{k \in g_v} Invoiced\ value_k^*\right)} \right\} \tag{2}$$

$$= \left| Invoiced\ value_i - Quantity_i \cdot UP_{Q2}(i) \right| \cdot \underset{over\ v=1-5}{\text{maximum}} \{ R_v(i) \} \tag{3}$$

where $$R_v(i) = \frac{1}{\sum_{k \in g_v} Invoiced\ Value_k^*} \cdot \frac{1}{O_v} \cdot f^{\,{}^{10}\log\left(\sum_{k \in g_v} Invoiced\ Value_k^*\right)} \tag{4}$$

Here we have taken the maximum over the five classifications, *v*=1–5. The sum of impacts for the five classifications might be a more relevant function.

Our choice of values for size effect (f) and for relative importance of aggregation variables (Ov) has made 6-digit CN, 3-digit SITC and 2-digit SITC approximately equally important when we judge the impact of a potential error. This can be seen in table 1.

**Table 1**
**The level of aggregation (*v* ) on which the reported lines had the greatest potential impact. Data from reference month October 2004. The size-parameter *f*=4**

| Level of aggregation (v ) | Importance of classifycation variable Ov | Number of observations flagged | Number of observations not flagged | Number of lines not possible to check | Total |
|---|---|---|---|---|---|
| Other CN8 | ---- | 0 | 0 | 0 | 0 |
| Important CN8 | 1.5 | 13 | 4,871 | 0 | 4,884 |
| CN6 | 1.0 | 600 | 145,761 | 59 | 146,420 |
| SITC3 | 0.3 | 488 | 137,340 | 17 | 137,845 |
| SITC2 | 0.2 | 390 | 90,605 | 12 | 91,007 |
| Arrivals/Dispatches | 0.1 | 0 | 0 | 0 | 0 |

# 5 Suspicion for errors in invoiced value and quantity

We soon learned that a symmetric search for errors in the variables *Invoiced value* and *Quantity* results in many more errors found for *Quantity* than for *Invoiced value* (see table 2 in section 15). As the primary objective of the statistics is to compute the summed value of foreign trade, we would like to prioritize the search for errors in *Invoiced value*. The unit prices bear no information on which of the two variables is in error. We therefore have constructed a rough indicator telling which of the two variables is most likely to be erroneous.

First we compute a measure of suspicion for the *Invoiced value* as such to be erroneous, i.e. without taking the *Quantity* into account. We do this by comparing the observed value to the quartiles, based on historical data, in the same way as in (1). Here it is important to include the variable Enterprise in the definition of the homogenous groups[4]. The same method is applied to the variable *Quantity.*

We then divide the suspicion for *Invoiced value* with the suspicion for *Quantity*. If suspicion for *Quantity* is zero, which it is for about half of the observations, it is replaced by a small number. Visually we have found that the following expression is linearly correlated to the proportion of errors in *Invoiced value*. This is our measure of suspicion for *Invoiced value* over *Quantity*.

$$Suspicion\_Va_i = \begin{cases} 1 & \begin{array}{l} if\ Suspicion(Invoiced\ value_i) \leq \\ Suspicion(Quantity_i) \end{array} \\ 1 + \log\left(\dfrac{Suspicion(Invoiced\ value_i)}{Suspicion(Quantity_i)}\right) & otherwise \end{cases} \qquad (5)$$

[4] Enterprises can chose between different ways to deliver information to Statistics Sweden. For one method each row on an invoice makes one observation, for another method data for one month are aggregated to a sum for in-/out-flow, country and 8 digit CN-code.

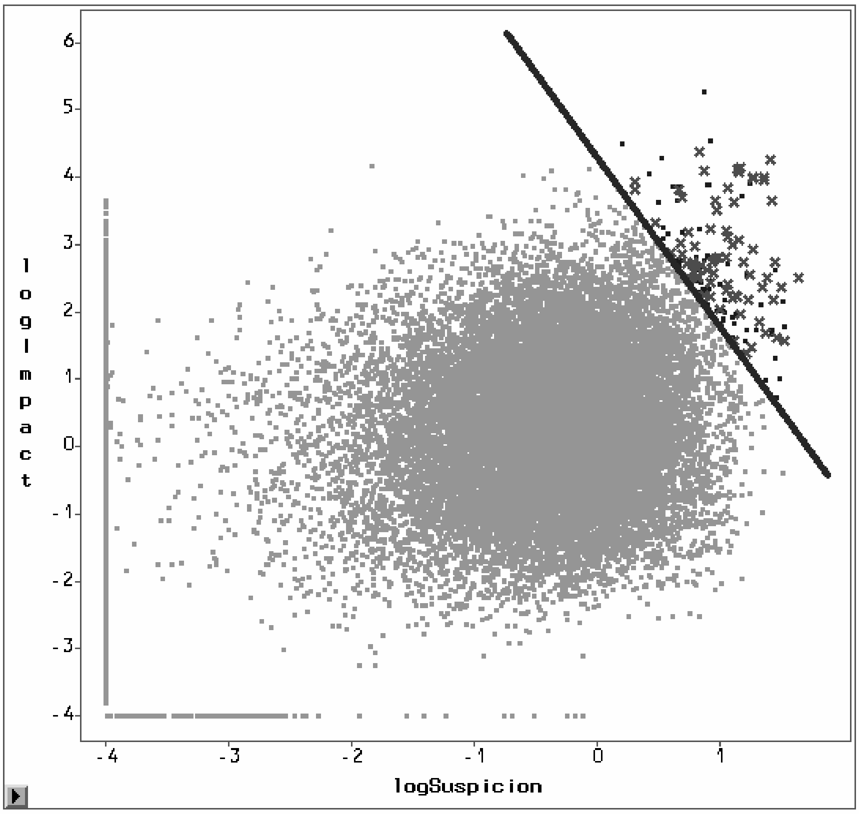
# 6 Score

We compute the score as a weighted geometric mean of the three variables we now have defined in (1), (2) and (5):

$$Score_i = Suspicion_i \cdot \cdot \left(Potential\ impact_i\right)^{P_{Imp}} \cdot \cdot \left(Suspicion\_Va_i\right)^{P_{Susp\_Va}} \quad (6)$$

In figure 1 it is illustrated that the boundary of the acceptance region is a line in the log-scale for *Suspicion* and *Potential Impact*, when *P_Susp_Va*=0. 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 to be erroneous.

**Figure 1**
**Acceptance region**



# 7 Study

Statistics Sweden has saved raw and corrected data for all months since the year 2000. We have used subsets of these data for searching for the best specification parameter values for our editing system. First we used data for seven months 2003 to make a rough limitation of the value ranges. Then we used 12 month of data for 2002 to search for the best values within a more limited set.

More often there are errors in the variable *Quantity* than in *Invoiced value*. Furthermore the errors are of larger magnitude for *Quantity*. In the study we omitted the observations where one or both of the variables were corrected with a factor <1/10,000 or increased with a factor >10,000.

**Table 2**
**Proportions of errors and impact for the variables *Quantity* and *Invoiced value***

| Variable | Proportion of data with errors (%) | Proportion of total impact of errors (%) |
|---|---|---|
| Quantity | 32.9 | 99.5 |
| Invoiced value | 7.7 | 0.5 |
| One or the other or both | 38.2 | 100.0 |

A question left open for discussion is whether we will be misled when we try to construct a new editing method by analysis of data that has been flagged by the old method. Is our conclusion concerning the optimum parameter values dependent on the method that has been used to flag the observations? Despite these concerns the method was implemented in January 2004, guided by the initial results of this study. Extensive process data are now produced every month in a continuous search for best parameter values.

We have used the following indicators to evaluate the performance of the editing method for different parameter values:

1) Imp(*Invoiced value*) = Maximum impact on published statistics due to errors in the variable *Invoiced value*
2) Imp(*Invoiced value*, *Quantity*) = Maximum impact on published statistics due to errors in the variable *Invoiced value* and/or the errors in the variable *Quantity*
3) Imp(*Invoiced value*, *Quantity*/100) = Maximum impact on published statistics due to errors in the variable *Invoiced value* and/or one hundredth of the errors in the variable *Quantity*[5]
4) Diff(*Invoiced value*, *Quantity*) = Corrections on the variables *Invoiced value* and *Quantity*, transformed to Value
5) Hit rate(*Invoiced value*) = Hit rate for any error in *Invoiced value*.

[5] Errors in *Quantity* are much more frequent and larger in size than errors in *Invoiced Value*. Division of the *Quantity* errors makes them comparable in this aspect.

6) Hit rate(*Invoiced value*, *Quantity*) = Hit rate for any error in *Invoiced value* and *Quantity*

## Step 1. Data for computation of quartiles

### Current data or historical data

A primary question in designing an editing method is whether current data or historical data should be used for computing the quartiles, including the median, that are used in the measures of suspicion and potential impact. For several product groups there are seasonal variations in both quantities and unit prices. Freshness of the data is therefore a good characteristic. Advantages of historical data are primarily that we can use as many objects as we may wish and secondary that the data are verified.

We have tested current unedited monthly data as well as one, two and three years of old data, i.e. 12–36 times as many observations. Here it is very significant that 1-3 years of monthly data are required to construct an efficient editing method. The study suggests using two years of monthly data.

### Grouping and number of historical observations

On which group do we compute quartiles in order to get both relevant and accurate measures of mean and dispersion for each observation *i*? Relevance depends on the homogeneity of the group, whereas accuracy depends both on homogeneity and number of observations. Historical data give, by nature, less relevant but more accurate estimates than current data.

We start off with the 6-digit CN-groups as the most heterogeneous groups we can ever accept. Then there are several possible variables to use for splitting these into more homogenous groups – with less objects in each. We have:

- 8-digit CN-code
- In/Out flow
- Enterprise
- Country (from which Sweden imports or to which Sweden exports)
- Last twelve months (in case several years of historical data are used)

If we split data into a cross-classification using all these variables, there will be none or only a few observations in many cells. Therefore we fix a minimum number of observations for the computation of quartiles, N_Obs. For a given order of the splitting variables we split each CN6-group hierarchically as long as there are more than N_Obs observations in the data. The priority order of the five splitting variables is also to be decided.

Choosing the one variable that gives the most homogenous groups is much the same problem as finding the variable that has the highest partial F-value in an analysis of variance with unit price as dependent variable and CN6 already in the model. Here we set a restriction that when there are fewer than N_Obs observations in a group the within sum of squares is computed around the average on the CN6-level.

**Table 3**
**Degree of explanation ( per cent) of total variation for the variable logarithm of unit price**

| Splitting variable | Minimum number of observations (N_Obs) | | |
|---|---|---|---|
| | 5 | 10 | 50 |
| Enterprise | 49.3 | 41.9 | 17.9 |
| Country | 9.0 | 8.3 | 5.4 |
| 8-digit CN-code | 5.6 | 5.5 | 4.8 |
| In/Out-flow | 4.0 | 4.0 | 3.6 |

The table indicates that editing should be made individually for each enterprise, if we have a lot of historical data. When we demand 50 instead of five observations for the computation, the power of this variable has declined most because there are so many different values. Beside this variable there seem to be little gain in splitting a 6-digit CN group.

There are interactions between the three parameters N_Obs, Priority_order and number of years of monthly data. If we have a large number of historical data, say three years, we can use a higher limit for the number of observations needed in the computations, N_Obs. The splitting variable Enterprise takes 16,000 values while the variable arrivals/dispatches only takes two, and therefore the second works equally well with a small and a high N_Obs.

For all the six indicators the best value of N_Obs is very clearly 2, which is an extreme value. We tried values from 2 to 64. Our conclusion is that it is important to split the population into as homogenous groups as possible. We have considered the risk that a systematic erroneous behaviour by enterprises can slip into the data if only two historical observations are needed per enterprise for computation of quartiles. We therefore suggest a value for N_obs in the range 4–8.

A priority order starting with Enterprise and Country might lead to a high hit rates and high total impacts in the search for errors, but for several combinations of other parameter values it has also lead to bad performance of the method. Setting the variables in the following order gives robust and good results[6]: 8-digit CN-code, In/Out-flow, Enterprise, Year and Country. We have also tested, with this order, if the last variable Country contribute to the potential impact of the flagged observations, and it did.

## Weighted or un-weighted quartiles

The historical data that will be used for computing quartiles for unit prices are very skewed on the variable *Quantity*. We will certainly get different results if we weight by *Quantity* or not. In some cases, especially if we use current un-edited data where there are errors, one or a few observations can influence the results substantially. For this reason we try a truncated *Quantity* as a weighting variable. The question of weighting interact with the choice of N_Obs. With very few observations in a group it might be better to compute quartiles with no weighting.

---

[6] All orders have not been tested.

For the Hit rate indicators and the pure difference indicator Diff(*Invoiced value*, *Quantity*) the best choice is not to weight. For hit rates this is understandable since all errors contribute equally much. When we want to detect errors that have a large impact on the summed values the best choice is to weight by *Quantity*, and preferable truncated for the largest values.

## Step 2. Definition of suspicion and potential impact

### Detail of grouping

When there are enough data to compute quartiles for several or all levels of groupings there is also a question if only the quartiles for the most detailed grouping should be used or if there is valuable information on more aggregated levels. Beside defining Suspicion and Potential impact with the most detailed splitting we have also computed them as averages over all possible splittings. The study did not yield a clear recommendation. Potential impact can preferable be computed with the most detailed splitting, while Suspicion rather should be computed as a mean of all possible splitting.

### Addition when the quartile distance is zero

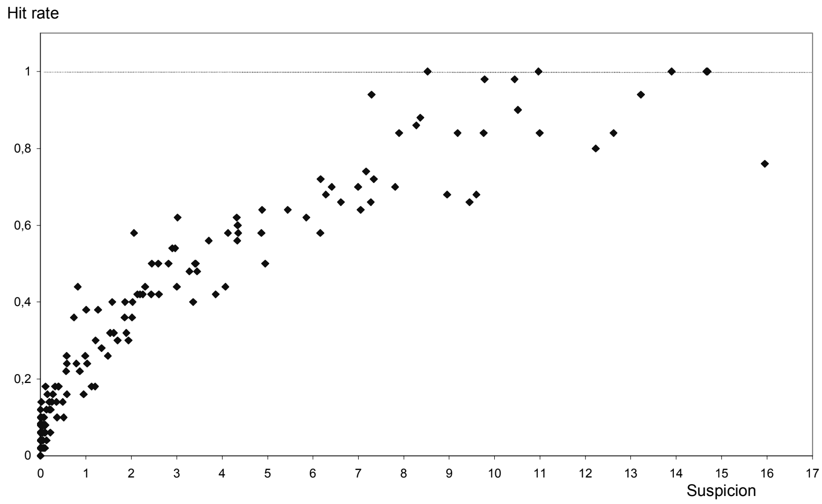
Hidiroglou and Berthelot propose that a fraction of the median replaces the quartile distance when the quartile distance is smaller than this fraction of the median. They propose the fraction 0.05. Since we analyze the logarithms of the unit prices, a one unit difference at any level on this scale means the same relative difference in unit prices. This would imply that the quartile distance should be replaced by a constant rather than by a fraction of the median. We tried several alternatives but found that 5-30 per cent of the median is the best.

### Suspicion versus probability of error

We want to pick out the set of observations that have the largest deviations between raw and corrected *Invoiced values*, valued by our formula (2).

Let us look at this problem from a probabilistic point of view. Assume we know the probability of error and the size of the potential error. Then we want to maximize the sum of products of probability of error and the potential impacts. Ideally we therefore want our variable Suspicion to be correlated with probability of error. The diagram shows that this is not the case.

**Figure 2**
**Hit rate versus Suspicion**




Footnote: Data have been sorted by Suspicion and aggregated into groups of 50 observations. Hit rate was computed as the proportion among the 50 observations in each group that had corrected data.

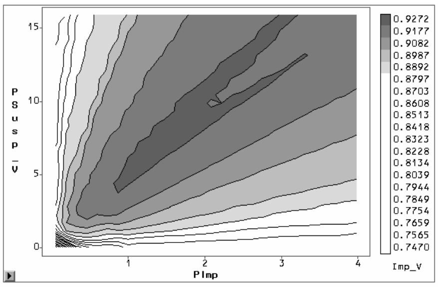
We have tested a few transformations of Suspicion. The simple transformation that Suspicion is unchanged up to 4, but the excess over 4 is reduced to 20 per cent, has worked well. This is especially so for the potential impact-indicators. For the Hit rate indicators a transformation should not be done, which is understandable as this measure is based on hit only – not the size of error.

## Step 3. Relative weights for suspicion and potential impact

The best choice of the relative weights $P^{Imp}$ and $P^{Susp\_Va}$ highly depends on the choice of indicator. If we want high hit rates, no or little weight should be laid on Potential impact, but if we want to detect errors in *Invoiced value* with high impact $P^{Susp\_Va}$ should be high.

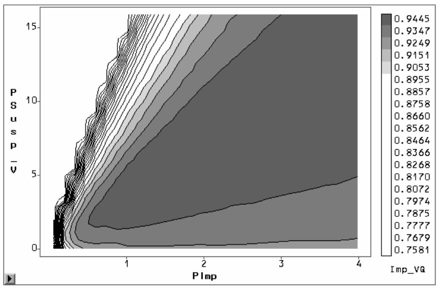
Imp (*Invoiced value*)

If we want to find observations with as large impact on the variable *Invoiced value* as possible we should select parameter values for $P^{Imp}$ and $P^{Susp\_Va}$ as for example 1 and 4 or 2 and 8..



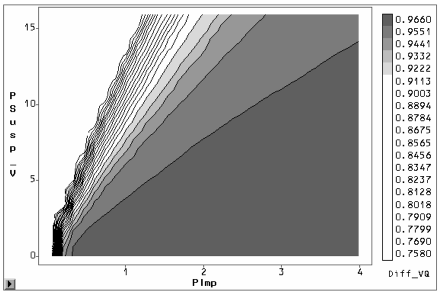

Imp (*Invoiced value*, *Quantity*/100)

If we want to get large impacts of the editing of both the variable *Invoiced value* and the variable *Quantity*, there seems to be a larger tolerance for different parameter values. $P^{Imp}$ and $P^{Susp\_Va}$ should be for example 1 and 2-4 or 2 and 4-8.



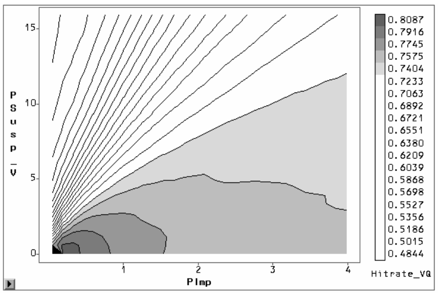

*Diff (*Invoiced value, Quantity*)*

*If we want to find summed corrections we should not be concerned for the suspicion for errors in* Invoiced value.



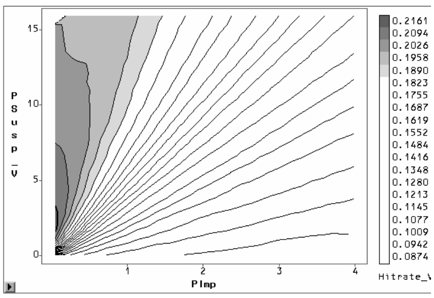

Hit rate (*Invoiced value*, *Quantity*)

The best parameter values to choose for getting highest overall hit rates are $P^{Imp}$ around 0.3 and $P^{Susp\_Va}$ to zero.




Hit rate (*Invoiced value*)

If we want high hit rates for the variable *Invoiced value* we should consequently disregard the potential impact and set the parameter $P^{Susp\_Va}$ to 3.



# 8 Implementation and results

The method was first tested for the reference month December 2003 and then implemented for the reference month January 2004.

## An embedded experiment in December 2003

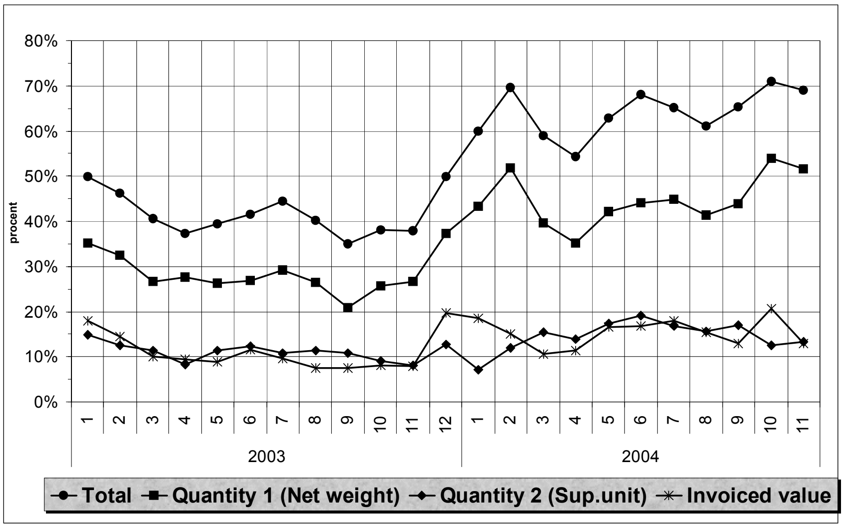
We tested the new proposed method for unit price checking as an embedded experiment in the ordinary production process for December 2003. The data were checked using both the old and the new method. The old method, which was not based on a score function, normally flagged about 2,400 observations each month. In the test the old method was modified so that only 1788 observations were flagged. The new method was set to flag 1,000 observations. It was found that 384 of these observations were flagged by both methods. The hit-rate for the observations flagged by the old method was found to be 39 per cent and the hit rate of the new method was 65 per cent. The sum of the corrections in the variable *Invoiced value* was 494 million SEK for the old method and 819 million SEK for the new method. Corrections of 29 million SEK was found by the old method but not by the new method. On the other hand the new method found corrections of 354 million SEK that the old method didn't find.

## Experiences from the production in 2004

Since the result of the test was satisfactory the new method was implementted in production for the reference month of January 2004. Due to needs for cost-cutting the number of edited lines had to be decreased from 2,400 observations to 1,500 observations per month.

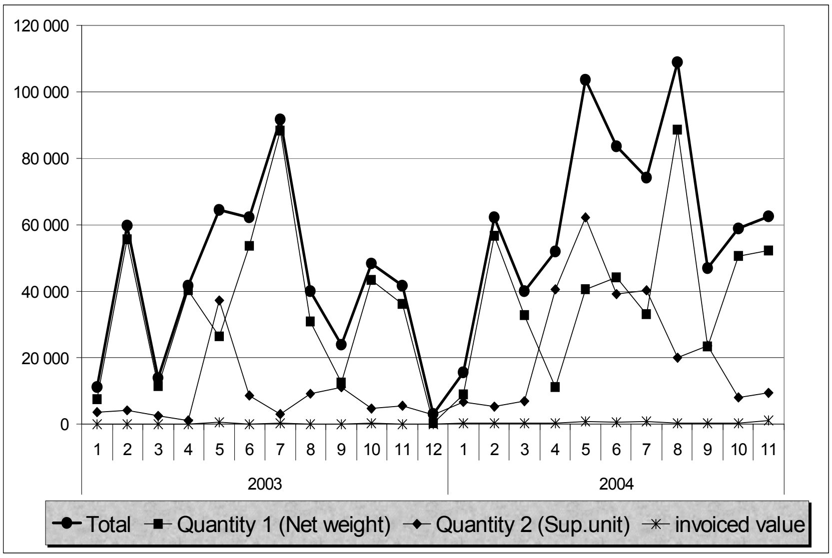
When the new method was introduced the hit rate increased from about 40 per cent to about 65 per cent (according to indicator 6). This can be seen in figure 3. The figure shows the total hit rate and the hit rate for the two quantity variables as well as for the variable *Invoiced value*. As can be seen most of the errors are found on the quantity variable *Net Weight*.

**Figure 3**
**Total hit rate and hit rate by variable**




Despite the decrease in the number of edited observations the impact seems to have increased somewhat as from January 2004. This is illustrated in Figure 4. The impact is here measured by the impact indicator 2. The figure displays the total impact and the impact by variable. Here it is very clear that the two Quantity-variables have the largest errors.

**Figure 4**
**Total impact of verified errors on the statistics and impact by variable**

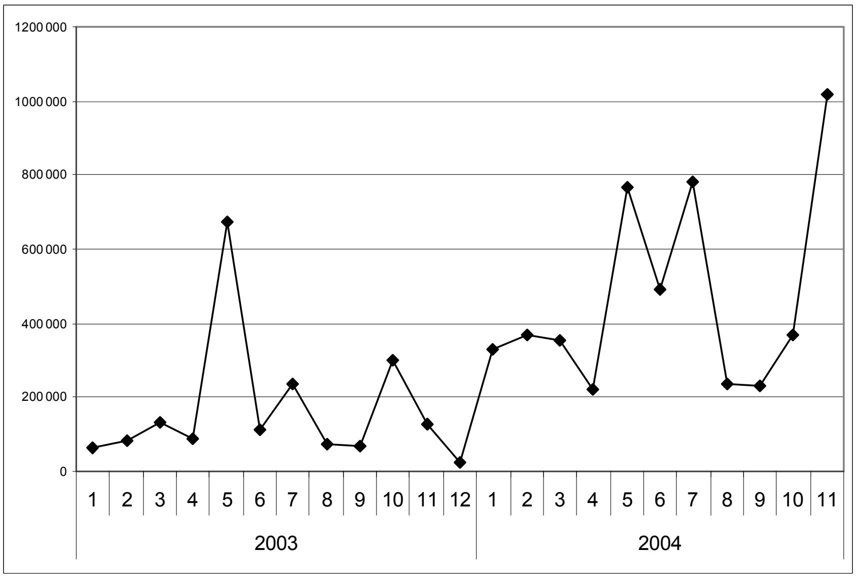



In figure 5 the impact on the variable *invoiced value* is displayed without the other variables (indicator 1). It is evident that the impact on *Invoiced value* has increased with the introduction of the new method.

We have changed two of the ten parameters from time to time (N_Obs and $P^{Imp}$) to see if this has an effect on the hit rate and/or on the impact of the errors found. The magnitude of errors in the data varies much between months, making it difficult to say what parameter values are the best without long series of process data.

For the first ten months 2004 the parameter $P^{Susp\_Va}$ is set to zero (Suspicion of *Invoiced value* relative to suspicion of *Quantity* is not considered). For the reference month November 2004 an embedded experiment was conducted to see if it was possible to increase the potential impact and hit rate for the variable *Invoiced value* by including the *Suspicion_Va* in the *Score*. The data were checked with and without the *Suspicion_Va*.

**Figure 5**
**The impact on *Invoiced value***




By each method 880 observations were flagged. It was found that 602 observations where flagged by both methods. The results concerning the hit rate are displayed in table 4 below. As can be seen the overall hit rate is somewhat lower for the extended method and the hit rate for one of the quantity-variables is considerably lower. The interesting point however is that the hit rate for the variable *Invoiced value* doubled from 8 per cent to 16 per cent by including the *Suspicion_Va* in the *Score*.

**Table 4**
**Hit rates for the experiment for reference month November 2004.**
**The extended method uses Suspicion_Va (suspicion for Invoiced value relative to suspicion for Quantity) while the ordinary method does not**

| | Ordinary method (without *Suspicion_Va)* | Extended method (with *Suspicion_Va)* |
|---|---|---|
| Overall | 76% | 73% |
| *Quantity 1* (Net weight) | 61% | 54% |
| *Quantity 2* (Sup. unit) | 13% | 12% |
| *Invoiced value* | 8% | 16% |

The total impact of errors found for the extended method is larger than for the ordinary method but the impact on *Invoiced value* is in fact smaller for the extended method. This is surprising. The extended method has found many small errors in *Invoiced value*, but it missed one large erroneous observation. We think it is an accidental occurrence, supported by figure 5 which shows the best result for November 2004.

**Table 5**
**Total value of corrections in Invoiced value and the impact measure for the three variables**

| | Difference in thousand SEK | Impact for *Invoiced value* | Impact for *Quantity* 1 (Net weight) | Impact for *Quantity* 2 (Sup. unit) | Total impact |
|---|---|---|---|---|---|
| Ordinary method | 2 558 402 | 1 692 | 1 356 810 | 3 791 181 | 5 149 683 |
| Extended Method | 2 488 222 | 1 582 | 1 355 935 | 3 795 959 | 5 153 476 |
| Difference | 70 180 | 110 | 875 | -4 778 | -3 793 |

In figure 6a can be seen the logarithm of the raw unedited *Invoiced values* plotted against the logarithm of the edited *Invoiced value*. The line from the origin with a 45 per cent slope is made up of observations with unchanged *Invoiced values*. The two lines parallel to this line consists of observations where the unedited values were 10 times too large or 10 times too small.

The corresponding picture for net weight is shown in figure 6b. Weights that are ten or thousand times too low or too high seems to be most common but weights that are 100 times too low or too high also exist.

**Figure 6a**
**Raw Invoiced value plotted against the edited Invoiced value**

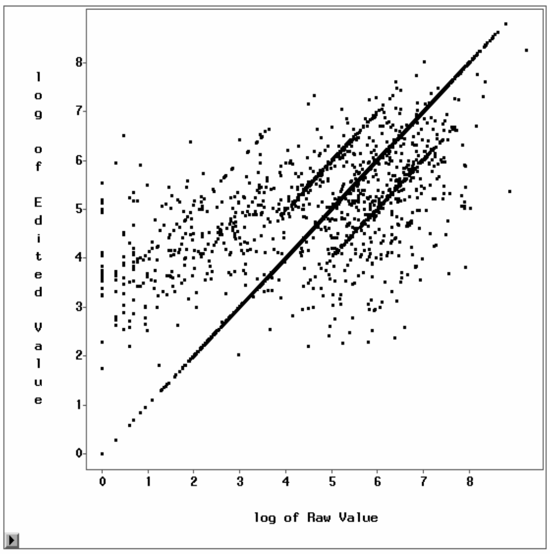
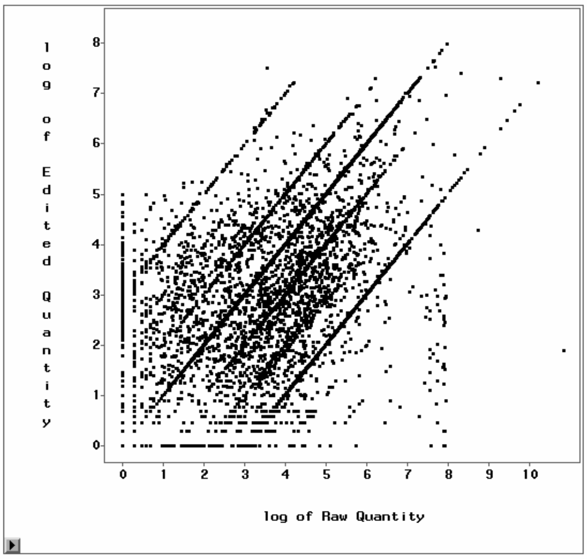



**Figure 6b**
**Raw net weight plotted against the edited net weight**



# Reference

Hidiroglou, M. A., and J.-M. Berthelot (1986), "Statistical Editing and Imputation for Periodic Business Surveys," *Survey Methodology*, Vol. 12, No.1, June 1986, pp. 73-83.

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

2005:08 Increased Automation of the Validation and Correction Processes in the Swedish Intrastat Production

2005:09 The Growth Rate Method for Production of Rapid Estimates of the Swedish Foreign Trade – Development, Evaluation and Preparations for the Implementation

2005:10 Svenskt producentprisindex (PPI) 1975–2004. En analys av tidsseriens integrationsgrad och säsongsmönster

2005:11 Kvaliteten på Sveriges statistik över import och export – En normativ och kvantitativ studie

2005:12 Implementing Simplified Intrastat Reporting in Statistics Sweden - Evaluation and Preparations for a Future Implementation

2005:13 FoU-verksamhet i kommuner och landsting – en pilotstudie inför kommande undersökningar

2006:01 Den svenska regionala tillväxten av i dag – en undersökning med utgångspunkt i "Den Nya Ekonomin"

2006:02 En flashestimator för den privata konsumtionen i Sverige med hjälpvariablerna HIP och detaljhandeln – En tidsserieanalys med hjälp av statistikprogrammet TRAMO

ISSN 1650-9447
ISBN 91-618-1321-4
ISBN 978-91-618-1321-6

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

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