

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Prague, Czech Republic, 14-17 October 1997)

Item 3 of the provisional agenda

**NATIONAL REPORT: INTRODUCTION OF AUTOMATIC ERROR SEARCHING
PROCESS ON FOREIGN TRADE DATA**

Submitted by Statistics Denmark¹

¹ Prepared by Tue M. Mathiasen.

1. INTRODUCTION AND MAIN RESULTS.....	2
2. STEPWISE PROCESS DESCRIPTION	3
2.1 KNOWLEDGE ON ABNORMAL DATA.....	3
2.2 HISTORIC DATA.....	3
2.3 TESTS PERFORMED ON NEW TRANSACTIONS.....	5
3. CONFRONTING THE DATA PROVIDER	8
4. PROCESS FLOW DIAGRAM.....	9

1. Introduction and main results

History

For several years the non-absolute error searching process on foreign trade data (INTRASTAT; EU-trade and EXTRASTAT; 3rd country trade) within Statistics Denmark has relied on a top to bottom manual search starting on the level of flow, commodity code and country. Some smaller transactions with goods not falling into ‘sensitive groups’ have been automatically searched through and compared to price/quantity-data for the last twelve months. The transactions found to be erroneous have had their quantity information automatically changed to fall into line with most recent twelve months data.

Reason to change the process

Mainly for three reasons we decided to develop and implement a new process:

- Need to unify and thereby document the process of error searching.
- Need to have an efficient process with minimal manual labour input.
- Need to ‘educate’ the responders (too many automatic corrections)

Main features of new process

We had the following requirements to the new process:

- No supposed errors should be rectified without confronting the data provider.
- Human labour should only be put into work, where data on a statistical basis seems implausible.
- The process must make sure, that those data having greater impact on statistics output are checked first - this is not an absolute criteria, because the structure of output changes over time.

Main results

A system has been developed, in which the search for errors is performed automatically by the means of a well documented statistical procedure based on mean and standard deviation calculations modified to exclude the use erroneous control data. The statistical procedure has a variety of possibilities to focus on certain types of errors as well as possibilities to change the significance level of the tests and thereby controlling the number of possible errors for further processing.

Finally a system to confront the data providers - currently being implemented in Statistics Denmark - is shortly described.

2. Stepwise process description

2.1 Knowledge on abnormal data

Accept master

To be able to take into account knowledge on abnormal data, an accept master is defined. On all entry combinations of flow, company identification, commodity code and country, it is possible - manually - to define minimum and/or maximum unit-prices (price/weight (pw) and price/supplementary unit (ps)) and weight/supplementary unit (ws) relationships (possible ranges include]0;∞]. Possible entry combinations are;

- Flow, company identification, commodity code, country
- Flow, company identification, commodity code
- Flow, company identification
- Flow, commodity code, country
- Flow, commodity code

This kind of information has been collected from various firms for several years by the error searching personnel. In general the information reflects abnormalities. The accept master is used later on.

2.2 Historic data

Control master

A control master is defined on transactions from most recent twelve months. Excluded are transactions which match entry combinations on the accept master.

The control master stores information on pw, ps and ws. The levels of calculation are flow/commodity/country (level 1) and flow/commodity (level 2). Information on level 1 is only being calculated if more than 9 transactions are recorded for the last year. The information calculated is mean (μ) and standard deviation (σ). In the system there are six possible ways to calculate the two statistics:

Type 0: All transactions are being used to calculate non-weighted mean and standard deviation μ_0 and σ_0 .

Type 1: All transaction are being sorted according to pw, ps and ws and the mid 80% transactions are being used to calculate non-weighted mean and standard deviation μ_1 and σ_1 . μ_1 and σ_1 are preferred to μ_0 and σ_0 if;

$$\sigma_1 \cdot 2 < \sigma_0$$

Tests have shown, that this procedure helps disregarding erroneous transactions (if there are relatively few of them) and thereby improving the use of the control master.

Type 2: In case of previous type 0, and if;

$$2 \cdot \sigma_0 > \mu_0$$

then;

$$\sigma_2 \equiv 0,45 \cdot \mu_0$$

Tests have shown, that this procedure helps diminish the effects of erroneous transactions (if there are relatively many of them) and thereby - again - improving the use of the control master.

Type 3: In case of previous type 1, and if;

$$2 \cdot \sigma_1 > \mu_1$$

then;

$$\sigma_3 \equiv 0,45 \cdot \mu_1$$

Same effect as above!

Type 4: In case of previous type 2, value weighted mean and standard deviation μ_4 and σ_4 are calculated. μ_4 and σ_4 are preferred to μ_0 and σ_2 if;

$$2 \cdot \sigma_4 < \mu_4$$

Tests have shown, that this procedure helps diminish the effects of erroneous low value transactions (if there are relatively many of them) and thereby - again - improving the use of the control master.

Type 5: In case of previous type 3, value weighted mean and standard deviation μ_5 and σ_5 are calculated. μ_5 and σ_5 are preferred to μ_1 and σ_3 if;

$$2 \cdot \sigma_5 < \mu_5$$

Same effect as above!

The information on the control master can be extracted on-line with information on mean, standard deviation, no. of transactions and type of calculation for pw, ps (and ws).

A summary of the control master is given below by showing the number of entries by flow, level of detail (commodity/country or commodity) and type of calculation for mean and standard deviation.

Table 1

<i>TYPE</i>	<i>Imports</i>		<i>Exports</i>	
	<i>level 1</i>	<i>level 2</i>	<i>level 1</i>	<i>level 2</i>
price/weight:				
0	6.245	1.765	8.037	2.228
1	10.238	3.193	10.446	2.637
2	4.362	1.262	4.453	1.169
3	5.137	2.682	4.487	1.495
4	1.179	393	1.344	382
5	60	16	71	11
price/supplementary unit:				
no suppl. unit	19.648	6.744	21.896	5.635
0	2.122	399	2.027	596
1	2.837	809	2.391	690
2	1.176	447	1.254	398
3	981	734	733	419
4	436	171	527	180
5	21	7	10	4
weight/supplementary unit:				
no suppl. unit	19.648	6.744	21.896	5.635
0	2.433	578	2.384	688
1	2.952	1.000	2.610	838
2	918	314	924	291
3	871	511	662	340
4	381	156	350	122
5	18	8	12	8

2.3 Tests performed on new transactions

Micro search

For each new or revised transaction tested minimum and maximum values for pw, ps and ws are computed.

The following formula (which - assuming a normal distribution - gives a 95 pct. acceptance area for transaction values (tv) over 500.000 dkr) is used to define minimum and maximum values:

$$\max, \min = \mu \pm 2\sigma f, \text{ where } \left\{ \begin{array}{l} f = 1 \text{ for } tv \geq 500.000 \text{ dkr} \\ f = \frac{500.000}{tv} \text{ for } 50.000 \leq tv < 500.000 \text{ dkr} \\ f = 10 \text{ for } tv < 50.000 \text{ dkr} \end{array} \right.$$

This part of the test gives narrow acceptance area to high value transactions. In the case of ws; $f \equiv 1$.

For pw and ps:

If $\min < 0,0$;
 $\min \equiv 0,001 \cdot \mu$

This ensures, that e.g. model catches transactions with low value but very high on kilograms or suppl. units.

In case of pw, ps and ws:

If $\min > 0,9 \cdot \mu$;
 $\min \equiv 0,9 \cdot \mu$ and
 $\max \equiv 1,1 \cdot \mu$

This ensures that pw, ps and ws are allowed to vary a certain degree.

If there are less than 10 transactions on the master on level 1, no mean and standard deviation is calculated and the test moves automatically to level 2.

If either pw or ps or both fall outside the acceptance area, an error code on level 1 is given:

pw: 1
 ps: 2
 pw and ps: 3

If the test moves to level 2, the following error codes are available:

pw: A
 ps: B
 pw and ps: C
 pw: D (if less than 10 transactions on the master on level 2)
 ps: E (if less than 10 transactions on the master on level 2)
 pw and ps: F (if less than 10 transactions on the master on level 2)

‘Macro’ search

To give errors affecting the macro-level some possible extra attention, a second step follows.

The incoming transaction is added to the control master data for checking purpose only. New means on pw and ps are computed;

- if possible on level 1 (calc. type should be 0 or 1) otherwise
- if possible on level 2 (calc. type should be 0 or 1 and no. obs. ≥ 10)

(It should be noted that in case of calc. type 1, the control master is based on only 80 pct. of the original no. of transactions.)

If the new unit-prices fall outside the following minima and maxima (which - assuming a normal distribution - gives a 50 pct. acceptance area) based on the original control master statistics:

$$\max, \min = \mu \pm 0,68\sigma$$

following error codes are given:

pw: X
 ps: Y
 pw and ps: Z
 pw: Æ
 ps: Ø
 pw and ps: Å

X, Y and Z is being used if the transaction already had a coding 1-3 or A-F from the first test. Æ, Ø and Å are used if no prior coding.

Using the accept master

For all transactions given an error code at this point, the information on the accept master is now being taken into account.

If the transaction is OK according to the accept master, the error code is deleted.

An extra test

A final test is being carried out, which broadly speaking is intended to catch errors that are not detectable on the basis of unit-prices. The test is again on level 1 if possible otherwise on level 2.

On level 1 it is tested whether a cell in the current month is relatively large compared to 1/12 of the cell on the control master (one years transactions). If there are less than 10 transactions on the control master for a specific cell on level 1, level 2 information is used by dividing the totals with the number of countries involved). By relatively large, we currently mean more than 10 mio. dkr and 100 pct. difference.

All transactions in the possible erroneous cells are given the error code '+'.

An overview of the distribution of possible errors is given in table 2.

Table 2

<i>Error Code</i>	<i>Import</i>	<i>Export</i>
1	49,35	50,67
2	7,85	3,77
3	3,37	1,90
A	5,62	4,54
B	1,04	0,62
C	0,73	0,34
D	1,01	0,90
E	0,13	0,17
F	0,28	0,37
X	16,16	20,03
Y	1,80	1,26
Z	1,74	1,73
Æ	5,50	7,22
Ø	0,77	3,67
Å	0,15	0,17
+	4,50	2,63
in total	100,00	100,00

In total, possible errors accounts for 3,8 pct. of all import transactions and 3,2 pct. of all export transactions.

3. Confronting the data provider

Error corrections are being performed differently in the two foreign trade collection systems.

Processing the possible errors in EXTRASTAT..

EXTRASTAT transactions with error codes are sent back to the fiscal authorities. Transactions given the codes X to Å will not enter the published statistics before amendments have been made or their correctness has been confirmed.

..and in INTRASTAT

INTRASTAT transactions having been given an error code are processed according to the specific error code given. There are two ways of processing:

- Confronting the company with possible errors by mail (this is being done for the error codes 1-3, A-C and X-Å which accounts for approx. 95 pct. of all possible errors).
- Confronting the company with possible errors by phone (this is being done for the error codes D-F and +).

Text coding

The error codes used for mailing has been chosen on the basis of statistical quality and the possibility of explaining the possible error in an automatic text-coding.

It is seen that possible errors in the ws-relationship has not been taken into account in the previous error coding. This is due to the fact, that a possible error in ws in theory would always accompany a possible error in pw or ps or both. When performing the text coding though, information on ws is taken into account. Also taken into account is the CN-8 nomenclature demand for supplementary units and the INTRASTAT demand for kilograms. The following table shows the coding.

Table 3

<i>Text coding</i>	<i>Error coding</i>	<i>ws OK?</i>	<i>Suppl.</i>	<i>Kg</i>
Possible error in value	3, C, Z, Å	OK	YES	YES
Possible error in weight	1, A, X, Æ		YES	YES
Possible error in suppl. units	2, B, Y, Ø		YES	YES
Possible error in value, weight and suppl. units	3, C, Z, Å	NOT OK	YES	YES
Possible error in price, weight or both	1, A, X, Æ		NO	YES
Possible error in price, suppl. units or both	2, B, Y, Ø		YES	NO

Currently being implemented:

On a weekly basis, all new and revised transactions are put through the error searching model. Letters are being produced for new transactions only, and we have set up a system, where all error corrections are performed in spreadsheets containing all relevant information of the transactions and control master.

Possible errors are sorted in company order which gives the possibility to have a system where all companies have a contact person in Statistics Denmark in all matters concerning errors in the reporting of foreign trade.

Absolute errors are processed in a parallel way.

4. Process flow diagram

