

**UNITED NATIONS STATISTICAL COMMISSION and  
ECONOMIC COMMISSION FOR EUROPE**

**CONFERENCE OF EUROPEAN STATISTICIANS**

**Work Session on Statistical Data Editing**

(Ottawa, Canada, 16-18 May 2005)

Topic (i): Editing administrative data and combined data sources

**DIFFERENT USES OF ADMINISTRATIVE DATA IN EDITING, ENUMERATING AND  
SAMPLING IN SURVEYS**

**Supporting Paper**

Submitted by Statistics Denmark<sup>1</sup>

**Abstract**

The external trade statistic in Denmark is based on two basic data collection systems, one regarding trade with other EU countries (Intrastat) and one regarding trade with the rest of the world (Extrastat). Firstly, the basic features of the statistical system and the use of administrative data. Afterwards the existing imputation models in Intrastat will be described and evaluated, concentrating on the cases where administrative data are used. Finally, some general remarks and future projects will be presented.

**1. THE EXTERNAL TRADE STATISTICAL SYSTEM IN DENMARK**

The continuous dissemination of external trade statistics in Denmark began in 1836. In the following more than 150 years the basic structure of the statistical system was more or less unchanged. Data was based on administrative data from the Central Customs Authorities treatment of commodity movements across the border. This material constituted a full coverage count of all commodity trade between Denmark and all other countries in the world. Presumably, this was not significantly different from how it was done in many other countries.

In 1993 a significant development in the European Union, the opening of the single market with free movement of goods, capital and labour, completely changed the prerequisites for the external trade systems in the EU countries. The data source disappeared because all the registrations on the national borders between the Member States (MS) were removed. Data collection changed to a system based on direct reporting to the statistical authorities from selected companies according to provisions in the EU regulation on statistical coverage (Intrastat)<sup>2</sup>. With the introduction of the internal market in EU the non-response issue became extremely important, and the experiences has proven this especially in the sense of resources used and effects on quality and timeliness. Intrastat was born.

Intrastat was introduced in Denmark (and the rest of the MS's in EU at that time) in 1993, with the purpose of collecting data and disseminating statistics regarding trade with the other EU-countries on a monthly basis. The old system was based on the administrative documents from the Customs Authorities

---

<sup>1</sup> Mr Carsten Zornig is Head of Section in the External Trade Division in Statistics Denmark. Mr Marius Ejby Poulsen is Head of the External Trade Division in Statistics Denmark. Mr Peter Ottosen is Special adviser in the External Trade Division in Statistics Denmark.

<sup>2</sup> See EU Council and Parliament regulation no. 3330/1991 and Commission Regulation no. 1901/2000. Today a new regulation apply: Council and Parliament Regulation no. 638/2004 and Commission Regulation 1982/2004.

consisting of data on trade with countries outside the EU – this system is called Extrastat and is more or less unchanged today.

So the basic framework consists of two systems, Extrastat and Intrastat:

- Extrastat is based on administrative data from the Customs treatment of commodities arriving from or leaving to a country outside the EU. Data consist of flow, partner countries, commodities, quantity in kg/supplementary unit, value, etc. All transactions (except the smallest ones below a certain threshold) are registered and sent to Statistics Denmark.
- Basically Intrastat is based on monthly declarations from a panel of companies (PSI's; Providers of Statistical Information). Approximately 10.000 companies are obliged to report to Intrastat in Denmark. The companies are selected on basis of the size of their annual trade (minimum 1.6 mill DKR for arrivals and minimum 4.1 mill DKR for dispatches). Administrative V.A.T. data are used for the delimitation of the population of companies combined with Intrastat reports from previous years. The basic data reported consist in principle of the same variables as in Extrastat.

In the paper only imputation in Intrastat is included, because this is the system where almost all imputation and estimation is made. In Extrastat some minor imputations are made due to transaction thresholds (low value trade where commodity, partner country and procedure code is not to be reported).

The basic imputation in external trade statistics in Denmark is not made in the traditional way, where the primary data is directly imputed. Data in this respect can consist of everything from individual transactions to periodic summation of total trade for each combination of flow, partner country, commodity and nature of transaction - the format and structure of data is not stable. It is more suitable to think about the unit as the missing value of trade for a specific company in a specific month divided on partner countries and commodities on the most detailed level. In other words the value that is imputed is the individual cell in the commodity/partner country matrix.

In Intrastat imputation is necessary because of non-response and trade below thresholds excluding the smallest traders (measured according to the value of trade). Non-response both consists of total non-response and partial non-response. The two cases are treated slightly differently.

Table 1  
**Estimation elements in the Intrastat system**

	Arrivals				Dispatches			
	Threshold limits	Non-response	Threshold transactions	Total/Coverage	Threshold limits	Non-response	Threshold transactions	Total/Coverage
	million DKK							
1993	1 361	9 169	794	109 923	1 398	8 462	380	129 429
1994	1 459	10 630	796	125 741	1 311	9 143	413	136 911
1995	1 756	15 892	994	183 418	1 554	12 617	705	186 189
1996	1 814	13 787	1 005	184 906	1 520	12 643	456	194 010
1997	5 238	16 833	725	206 609	4 420	9 986	391	208 668
1998	5 579	21 229	866	219 709	4 500	9 602	472	209 959
1999	5 704	22 112	1 101	221 986	4 449	12 909	624	226 752
2000	5 716	20 705	1 156	249 972	4 671	14 648	648	266 962
2001	6 436	24 476	937	256 144	4 670	11 942	727	272 182
2002	6 858	31 629	808	277 569	4 760	18 976	491	285 464
2003	6 991	26 584	816	259 015	4 809	14 569	319	279 673
	pct.							
1993	1.2	8.3	0.7	98.8	1.1	6.5	0.3	98.9
1994	1.2	8.5	0.6	98.8	1.0	6.7	0.3	99.0
1995	1.0	8.7	0.5	99.0	0.8	6.8	0.4	99.2
1996	1.0	7.5	0.5	99.0	0.8	6.5	0.2	99.2
1997 <sup>1</sup>	2.5	8.1	0.4	97.5	2.1	4.8	0.2	97.9
1998	2.5	9.7	0.4	97.5	2.1	4.6	0.2	97.9
1999	2.6	10.0	0.5	97.4	2.0	5.7	0.3	98.0
2000	2.3	8.3	0.5	97.7	1.7	5.5	0.2	98.3
2001	2.5	9.6	0.4	97.5	1.7	4.4	0.3	98.3
2002	2.5	11.4	0.3	97.5	1.7	6.6	0.2	98.3
2003	2.7	10.3	0.3	97.3	1.7	5.2	0.1	98.3

<sup>1</sup> Prior to 1997 the threshold limits for imports and exports were DKK 800,000 and 500,000 respectively. From 1997 to 2004 the thresholds were DKK 1,500,000 and DKK 2,500,000. Finally, in connection with the adoption of the current EU Council and Parliament regulation for Intrastat (no 638/2004), the thresholds were raised to DKK 1,600,000 and DKK 4,100,000.  
Source: External Trade of Denmark 2003, Statistics Denmark 2004.

One of the major tasks in the production of the statistics is to minimize the share of non-response. As it appears in table 1 it has not been possible to succeed with this in a satisfactory way. For arrivals the non-response in 2003 was 10.3 (compared to 8.1 in 1997) and for dispatches 5.2 (4.8 in 1997). Adding to this there appears an element of 3.0 pct. estimation for arrivals due to threshold transactions and limits and 1.8 pct. for arrivals. In total the estimation elements are approximately 13 pct. for arrivals and 7 pct. for dispatches.

Non-response has an influence on the uncertainty of the statistics, which lead to larger revisions of the aggregated figures (annoying the economic analysts) and the detailed figures (annoying the market analysts). Consequently imputation has a significant importance in the objective of keeping the continuous revisions down. To illustrate the problem, figures on the revisions in Intrastat in the years 2000 to 2002 are shown in table 2.

Table 2  
Revisions of the Danish figures on EU-trade, 2000 to 2003

	Actual		Numerical	
	Mill. DKK.	Pct.	Mill. DKK.	Pct.
<b>2000</b>				
Import .....	3,659	1,5	8,727	3,5
Export .....	16,695	6,3	17,014	6,4
<b>2001</b>				
Import .....	-599	-0,2	6,827	2,7
Export .....	2,677	1,0	12,390	4,6
<b>2002</b>				
Import .....	2,237	0,8	21,826	8,0
Export .....	-751	-0,3	21,651	7,6

Note: *Actual revision* is calculated as the value of the latest version minus the value of the first version (e.g. the import from EU in December 2002 is published the first time in February 2002 and the last time in October 2002). The figure says something about the quality of the first estimate published – the Flash estimate.

*Numerical revision* is calculated as the actual revisions made up in numerical values (i.e. a revision of -500 mill. US\$ enters as +500 mill. US\$). The figure says something about the total volume of the revisions.

Source: Kvaliteten af udenrigshandelstallene, juni 2004, [www.dst.dk/Udenrigshandel](http://www.dst.dk/Udenrigshandel) (Quality report disseminated regularly on the internet, only published in Danish)

When looking at the actual revisions the table show that after a less successful year in 2000, the precision was significantly improved in 2001 and 2002, represented by a large decrease in the revisions. On the other hand the total volume of the revisions increased from 2001 to 2002. Detailed studies of the reasons for revisions have not been carried out, but although the capability of the imputation models and techniques are not the only sources of explanation, they most certainly have a dominant role.

The most detailed level of dissemination in the external trade statistics is

Month\*Flow\*Commodity\*Partner country\*Measurement unit

where Month = January, February, ..., December

Flow = Arrivals (imports) and dispatches (exports)

Commodity = 8 digit code based on the Combined Nomenclature CN (which is based on the Harmonised System HS)

Partner Country = 2 letters code according to international standards

Measurement Unit = Value (DKR), amount (kg) and supplementary unit

Data are stored as traditional data records with a specific sequence of the values of the variables reported. Normally the PSI's make one report for each month, and they can choose to report all individual transactions or report them in an aggregated form. There is a maximum level of aggregation (see the list above supplemented with type of transaction) but no strict rule below this level. Consequently as a data recipient we do not know if a given record is a single transaction or an aggregate of several similar transactions. Data are imputed without paying attention to the nature of transaction, i.e. the imputed trade is treated as an aggregate of several similar transactions.

So we are not in a situation where we have a definitive number of units in our population. This is contrary to the normal case in socio-demographic statistics where the number of people registered on a given time or the survey sample is known - normally data consist of one record per person and hence the total number of records is definitive. In the case of external trade, imputation is concerned with the reparation or consolidation of missing/erroneous values. In Intrastat several cases can appear:

1. Values in a given record can be missing or erroneous

2. Records (trade) can be missing (partial non response)
3. Trade can be missing (total non response)

Drawing the attention to the dissemination level, which in practice also determines the level of imputation, the relevant unit is the cell in the Partner country\*Commodity matrix, where a given cell consist of the value of trade for a specific combination of the two. In chapter 3 the filling out and consolidation of this matrix (or matrices because we actually have one for each month not yet finalised<sup>3</sup>), using imputation/estimation techniques is described.

In conclusion, regarding imputation in Intrastat the statistical unit is not solely the physical unit (the respondent being a firm or a company) but the cells in a Partner country/Commodity matrix. The challenge is to distribute the missing trade in this matrix.

## **2. USE OF ADMINISTRATIVE DATA**

### **2.1 The role of the Central Business register**

Basically the Intrastat population is based on data from the Central Business Register (CBR). This register contains a variety of administrative information defining and describing all business units in Denmark. Different levels of registration of business units exist and one of the great challenges in using the units for collection of statistics is to combine/merge the related units in an appropriate way, in order to get a comprehensive description of the company.

One of the key prerequisites for the establishment and effective maintenance of the CBR is a good and trustful cooperation with the administrative authorities in charge of collecting the administrative data. This is the case in Denmark and therefore the CBR is the back bone in almost all statistical data collections regarding the business sector - either as basis for sampling or as a direct source (e.g. main economic features and employment).

First of all CBR consist of the data elements necessary to identify the companies, which is important in connection with administration of the population (CBR-number, address, etc.). Another important data element is the activity code, which is still used in calculation of a distribution key for the imputation and distribution of trade.

### **2.2 V.A.T. data**

Apart from the administrative CBR the different administrative registers on business units cover overall subjects in the business sector. One of the subjects covered is Value Added Tax (V.A.T.), and in this connection two boxes on the V.A.T. form are of special interest to external trade statistics.<sup>4</sup>

Box A: Total EU buys

Box B: Total EU sales

These two boxes (together with a third box C including trade in services and non-EU trade) were introduced on the V.A.T. scheme in 1993 when Intrastat was started. Consequently each company have to report this information regularly.<sup>5</sup>

The Intrastat population is up-dated twice every year, and in this connection the information in box A and B are used to delimit the population. This information is supplemented with earlier reports from Intrastat.

<sup>3</sup> As an example, today (11 April 2005) the figures for December 2003 and earlier are all finalised (that is published in their final form), whereas the figures for January 2004 up to February 2005 still have a provisional status.

<sup>4</sup> Data from the V.A.T. scheme is in Statistics Denmark primarily used (including box A, B, and C) for statistics on turnover.

<sup>5</sup> Monthly for large companies, quarterly for medium-sized and half-yearly for small companies.

Each company with trade above the exemption thresholds are included in the population of PSI's - having to report to Intrastat on a monthly basis.

However, the main interest in this paper is the use of the data from box A and B for imputation, see chapter 3. Other administrative data from the business register environment, which are relevant in connection with estimation and imputation, are data on sector of activity and data from the V.A.T. Intra community Exchange System V.I.E.S. (also known as Intra Community Sales List). The latter is used for V.A.T. control purposes by Customs, and contains information on total dispatches divided by partner companies abroad – in chapter 4 the possible use of this source is described.

### **3. IMPUTATION OF EXTERNAL TRADE STATISTICS TODAY**

#### **3.1 General description of the models**

As described above, the Danish trade figures are divided into trade with EU-countries (Intrastat) and trade with non-EU-countries (Extrastat). Imputation of data is only carried out regarding Intrastat, as the Extrastat in principle makes up a complete coverage.

The coverage under Intrastat is approximately 90 per cent, see table 1, which partly is due to trade below the thresholds and partly due to non-response. As a consequence, the 10 per cent non-reported EU-trade must be estimated, which is carried out in two separate models:

- The Flash model, comprising the current months
- The Master model, comprising all revised months

In principle, the aim of both models is to estimate the extent of the non-reported EU-trade. In the Flash model, the amount of non-reported trade is calculated as the residual between estimates of the total imports/exports and the reported trade data, whereas in the Mester model, the non-reported trade is calculated for each individual enterprise and enumerated to a total amount of non-reported trade. Subsequently, the reported trade distribute the estimated trade over countries and commodities, which is in principle carried out similarly in both models. When the imputation is completed, full correspondence is achieved between the most detailed trade data and the aggregated trade figures.

#### *Flash*

The Flash model is not an imputation model but a linear regression model based on a panel of enterprises. The model is presented to give the complete picture of the estimation system.

The availability of V.A.T. data for imputation purposes sets a limit to the timeliness of data produced using a V.A.T. based imputation system. To further improve the timeliness of the external trade statistics a so-called Flash estimator was introduced in 2000.

This method estimates the first figures on Intra community trade. The principle in the method is to optimize the use of the information already available at the earliest possible time.

When the first figures are compiled starting approximately 35 days after the reference month approximately 60 pct. of the final trade has been reported. These reports cover a range of different kind of enterprises. Some have reported every month on time and correct the first time, some have never reported on time before and some report so badly that the report contains no useful information. The first group is the interesting one: consistently timely reporter with high quality reports.

These traders represent the estimation panel of enterprises selected from scratch each month based on the latest reported figures. For a PSI to be included in the panel it must fulfil three criteria:

- Timeliness: The PSI has reported import and/or export timely for the last 12 months.
- Correctness: The reported trade for the past 12 months is reported correct the first time – in practice, we allow for small changes below 5 per cent of the initial reported trade

- Validity of report concerning the reference month: The reported trade must have been relatively correct, in practice, the reported trade lies within the boundaries of 50 per cent the minimum and 200 per cent of the maximum reported trade in the last year. This means, for example that a company that sometimes have zero-reports (reporting no import or export) will be included in the panel with a zero-report, where as a company that has not reported zero before, will be excluded.

A different panel is selected for each flow.

For the panel the reported trade data are generated, as they appeared on the n'th day after the end of the reference month, i.e. if the panel has been constructed 30 days after the reference month, the historic trade data are generated as they appear on the 30' Th day after the respective reference day. The panel data consist of the n'th day reports for each month since January 1995 for the selected companies.

The times series with panel data are used in an Ordinary Least Square regression model. Ten different regressions are made: total arrivals, arrivals from Germany, UK, Sweden and the remaining EU countries, total dispatch, dispatch to Germany, UK, Sweden and the remaining EU countries<sup>6</sup>.

In the regression, additional variables are entered. For most of the regressions eleven seasonal dummies are entered. In some of the regressions, a significant break occurs in 1997.

The regression is as follows

$$T = a_0 + b_1P + b_2S + b_3D + e$$

where T=external trade (arrivals or dispatch)

P=the panel trade

S=the seasonally dummies to adjust for seasonal fluctuations

D=a set of dummies to take into account breaks and outliers

The estimate of missing trade is the difference between total estimate and the reported trade. The missing trade is spread out on commodities using a method similar to the one used in the Mester system with regard to complete non-response see description below.

As no imputation is carried out directly in this model, the rest of this section will only pay attention to the Mester model.

#### *Mester*

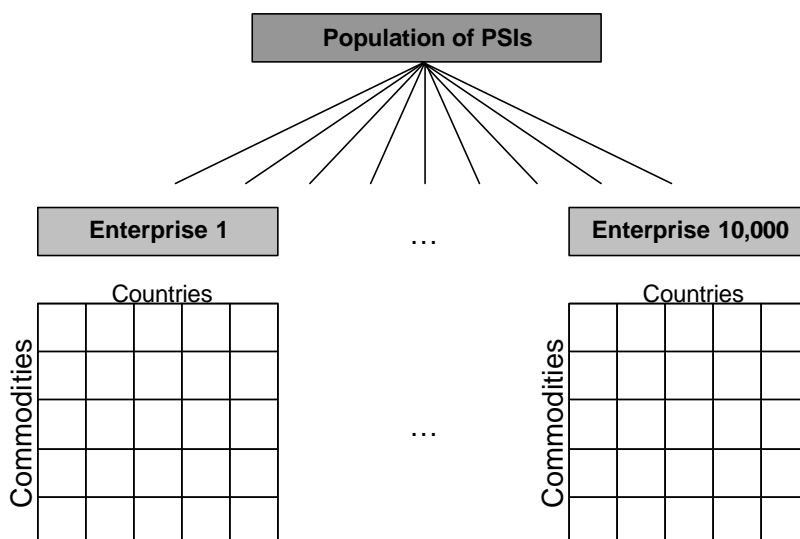
The Mester model comprises all months not comprised by the Flash model, i.e. the revised months.<sup>7</sup> The imputation is carried out for all enterprises, which are involved in external trade. The imputation is based on the administrative V.A.T. data, collected by the Danish Customs authorities.

Thus, the overall aim of the Mester model is to delimitate the enterprises affected by non-response, and then impute the missing trade distributed over countries and commodities. In principle, a matrix containing all combinations of commodities and countries must be established for all 10,000 enterprises. For each of those matrixes it must be decided, whether the reported trade is representative for the actual trade or not. If not, the matrices must be repaired by imputation, in order to eliminate non-response.

This can be illustrated as below.

<sup>6</sup> Trade for the 10 new EU member states (from May 2004) are estimated separately.

<sup>7</sup> An overall sketch of the Mester model is given in the end of chapter 3.



### 3.2 Data available for the Mester model

The data used in the Mester model is Intrastat as the source for the detailed data, and V.A.T. as the source for the aggregated enterprise level.

#### *Intrastat*

Intrastat was introduced in Denmark in 1993, together with the introduction of the internal EU market. Intrastat forms replaced the Single Administrative Document (SAD), which was used by the Customs authorities for clearance of goods crossing the external frontiers. Intrastat is a far more simple system compared to the SAD document, as only few data are collected.

Under Intrastat, the following import and export data are collected on a monthly basis for all enterprises:

- Nature of transaction (ordinary trade/returned goods/processing etc.)
- Commodity code according to EUs Combined Nomenclature (CN), comprising approximately 10,000 individual commodity codes. The 8-digit CN codes are derived from the 6-digit codes in the Harmonised System by adding a subdivision of two digits.
- Partner country (country of consignment for imports/country of final destination for exports)
- Value in DKK, weight in kilos and/or supplementary units (litres/square metres etc.)

In addition, the Intrastat system caused a reduction of the coverage, as small enterprises are excluded from the population of PSIs. According to the Intrastat regulation, the coverage of the enterprises comprised by the population of PSIs must make up 97 per cent of the total EU trade. Thus, only enterprises with an annual import exceeding 1.6 million DKK respectively an annual EU export exceeding 4.1 million DKK are comprised by the PSI population. This information is derived from the V.A.T. data, resulting in a population of approximately 10,000 enterprises (exempting approximately 30,000 enterprises from the obligation to report to Intrastat).

#### *V.A.T.*

Similar to Intrastat, a new V.A.T. form was introduced together with the internal EU market. Information is collected by Customs on a monthly, a quarterly or a half-yearly basis, depending on the annual turnover of the enterprises.

The new information available from the V.A.T. data was:

- Box A: total value (excl. V.A.T.) of commodities bought in other EU countries
- Box B: total value (excl. V.A.T.) of commodities sold to other EU countries. Enterprises having EU-exports are also obliged to provide information on the receiver of the commodities on a quarterly V.I.E.S. form.
- Box C: total value (excl. V.A.T.) of commodities and services exported. This information is not used in the present models.

Even though the information available from the V.A.T. form is rather limited, the information is very important, as all enterprises involved in EU trade are obliged to fill in the V.A.T. form. Thus, the coverage in principle is total. Therefore, the V.A.T. data can be used to enumerate the extent of the EU imports (from Box A) and the EU-exports (from Box B).

### 3.3 Preparation of data

Before data can be merged and used in the Mester model, some preparation is necessary, in order to achieve complete comparability between data.

#### *Intrastat*

Trade data collected under Intrastat are divided on the nature of transaction, and not all of these transactions are comparable to the V.A.T. data.

This comprises among others trade with goods for processing, returned goods etc.<sup>8</sup> Other acceptable discrepancies may also occur between Intrastat and V.A.T. This comprises among others periodicity, use of different enterprise units for Intrastat and V.A.T., etc.

These types of discrepancies can be explained due to conceptual differences between the two data sources, and consequently these transactions are treated separately in the model. Other conceptual discrepancies are detected during the error detection routines, and those discrepancies are rectified from the model as well.

#### *V.A.T.*

The V.A.T. figures are reported on a monthly, a quarterly or a half-yearly basis, and therefore the data available for the model will only be a complete set of data two times within a year (when the half-yearly data are reported).

This lack of data must be imputed, which is carried out in two steps. First, the monthly reports are completed during imputation and second, the quarterly and the half-yearly reports are completed and imputed in accordance with the distribution that can be emerged from the monthly reports.

The quantity of data that must be imputed depends on the amount of V.A.T. data available. Due to time lags, the monthly V.A.T. reports are not available for approximately one month, while the quarterly and the half-yearly reports not are available for respectively three and six months.

The basis of the imputation of the monthly V.A.T. data is the reports from enterprises having reported each month within the last 24 months. This delimitation is made, in order to avoid unexpected shifts due to new enterprise structures (new enterprises, mergers etc.). Based on those V.A.T. reports, the trend for respectively the import figures and the export figures is calculated. The calculation of the trend is based on the *average least square* model.

The idea of the least squares method is to find the best fitting curve to a given set of points by minimizing the sum of the squares of the residuals of the points from the curve. The sum of the squares of the offsets is used instead of the offset absolute values because this allows the residuals to be treated as a continuous differentiable quantity.

---

<sup>8</sup> An example of trade causing a conceptual discrepancy: Goods for processing will cause discrepancies between the two data sources. Goods imported for processing are included in Intrastat imports with the value of the good before processing and afterwards in the export statistics with the value of the good including the value of the processing. In the V.A.T. figures, and so the value reported in Box A will be zero, whereas the value reported in Box B will be the value of the processing.

This fitting technique is a simple form of linear regression, which solves the problem of finding the best fitting straight line through a set of points.

The linear line describing the trend can be formulated as follows:

$$y = \beta_0 + \beta_1 x \quad \text{where} \quad \beta_0 = y - \beta_1 \bar{x} \quad \beta_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

The trend calculated for this sample of enterprises is used as the basis of the imputation as well as the monthly, the quarterly and the half-yearly V.A.T. figures.

Based on the calculated trend, the monthly reports for the enterprises included in the trend calculation can be forecasted for months without V.A.T. The forecasted V.A.T. data are adjusted for seasonal fluctuations in accordance with the V.A.T. data for the corresponding month in the previous year. The result of this operation is a complete set of V.A.T. data for the sample of enterprises included in the trend calculation.

This procedure is carried out at the aggregated level, but is subsequently calculated at the enterprise level to ensure that data are available at the most detailed level.

The V.A.T. data for the remaining monthly reporters and for the quarterly and the half-yearly reporters are forecasted in accordance with the calculated trend. When data has been forecasted, the quarterly and half-yearly reports are re-distributed on months, in accordance with the distribution emerged from the monthly reporters. This re-distribution, which is carried out in order to maintain seasonal characteristics at the aggregated level, comprises as well the reported as the forecasted V.A.T. data.

The result of the imputed V.A.T. figures is a complete set of data, providing monthly information on the total EU trade for all enterprises.

### 3.4 Error checking routines

Three individual error checks are carried out, before data are accepted as comparable.

First, all Intrastat data are subject to several error checks. Among others, all major individual values in the Intrastat reports are checked manually. The aim of these checks is to rectify errors having impact at the aggregated trade data level.

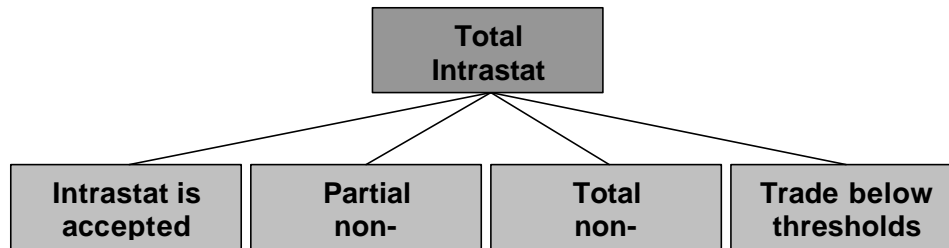
Second, the V.A.T. data are subject to a preliminary error-check, involving a manually check of all major reports (Box A and Box B) deviating from the expected value, i.e. all values outside the interval defined by the mean value plus/minus two times standard deviation. This check is carried out to rectify obvious V.A.T. errors, in order to reduce the main error source to the Intrastat data.

The third check consists of a manually investigation of all enterprises having major discrepancies between Intrastat and V.A.T. The check is an enterprise specific comparing of the Intrastat and the V.A.T. data. The aim of this error check is to identify and specify the enterprise specific non-response. The result of this error check is primarily input at the aggregated enterprise level, via information on the (expected) correct total Intrastat values for the enterprise and/or information on the correct V.A.T. values. If the Intrastat data are identified as incorrect, the enterprises are obliged to send in a new set of data. However, this procedure is very time-consuming, which is why the information on the correct total Intrastat value is very useful for the estimation procedures. Many discrepancies between Intrastat and V.A.T. can be explained as conceptual, due to periodicity, use of different enterprise units etc.

### 3.5 Enumeration to complete coverage

When Intrastat and V.A.T. data are comparable, data are ready to be merged, in order to estimate the extent of non-reported EU-trade in the Mester model.

The first step in the model is to divide the Intrastat population into four main groups, as indicated below. The aim of this division is to delimitate the set of enterprises affected by non-response, and to identify enterprises having trade below the thresholds.



#### *Intrastat is accepted*

This group includes enterprises within the population of PSIs, for which the Intrastat reports seem to be representative, i.e. non-response is assessed equal to zero and consequently no estimation is necessary. Generally, Intrastat is accepted, if the value reported to Intrastat exceeds the V.A.T. value. This assumption should be seen because of three factors:

- It is far more time-consuming reporting to Intrastat
- All major Intrastat values have been checked manually.
- Many V.A.T. errors have been detected and corrected during the preliminary check of data.

Thus, as all trade data have been reported, all cells in the country-commodity matrix are known, and no imputation is necessary.

#### *Partial non-response*

Partial non-response for enterprises within the PSI population is assumed, if the V.A.T. data exceeds the Intrastat data, and the discrepancy cannot be explained as conceptual.

In these cases, the extent of non-response is estimated, based on historical information. For imports, 43 per cent of the discrepancy is assigned to non-reported Intrastat, respectively 53 per cent for exports, whereas the rest is assigned to incorrect V.A.T. data. These percentages are derived from the experiences gained from the enterprise specific error checks. The estimated non-response constitutes the amount of trade, which is subsequently imputed in the enterprise specific country-commodity matrixes.

As some trade data has been reported, some information is available on the level of individual cells in the country-commodity matrix, so only the remaining cells is to be imputed.

#### *Total non-response and Trade below thresholds*

Total non-response and trade below the thresholds are in principle treated in the same way. The only difference between the two groups is the information, that enterprises trading below the thresholds not are included in the PSI population.

In these situations, the extent of non-reported trade (i.e. the amount of trade to be imputed in the country-commodity matrixes) is estimated to be equal to the value on the V.A.T. form, unless additional information is available from the enterprise specific error checks.

As no trade has been reported for enterprises within these two groups, no information on the individual cells in the country-commodity matrix is known.

### 3.6 Imputation of countries and commodities

The imputation of non-reported trade is carried out within the two groups:

- Partial non-response
- Total non-response/trade below the thresholds

In principle, the imputation of non-reported trade is carried out in a similar way for the two groups. It is common for both groups that the imputation of non-reported trade is based on the trade pattern that can be derived from previously reports. However, the two groups must be treated different due to the amount of information available from the reported Intrastat data.

#### *Partial non-response*

Enterprises within this group have been delimited due to the fact, that some Intrastat data have been reported. This reported trade is used to calculate the trade pattern for all enterprises in the group, based on the enterprise specific Intrastat reports within the previously 12 months. Those trade patterns are converted into distribution keys, which are assumed to represent the enterprise specific distribution over countries and commodities.

For all enterprises, the non-reported trade can now be distributed over countries and commodities, in accordance with the distribution keys.

#### *Total non-response/trade below the thresholds*

The situation for the enterprises within this group is more complex, as no information on the trade pattern is available from the Intrastat reports.

This is dealt with by substituting the enterprise specific trade with trade from similar enterprises. The assumption is, that the trade pattern for these enterprises on a large scale is similar to the trade characterised by enterprises within the same industry (indicated by the activity code from the CBR). Thus, for all industries, the trade pattern is calculated on basis of the trade for the previous 12 months.

Similar to the group of partial non-response, the trade pattern is converted into distribution keys, which are assumed to represent the distribution over countries and commodities for the industry.

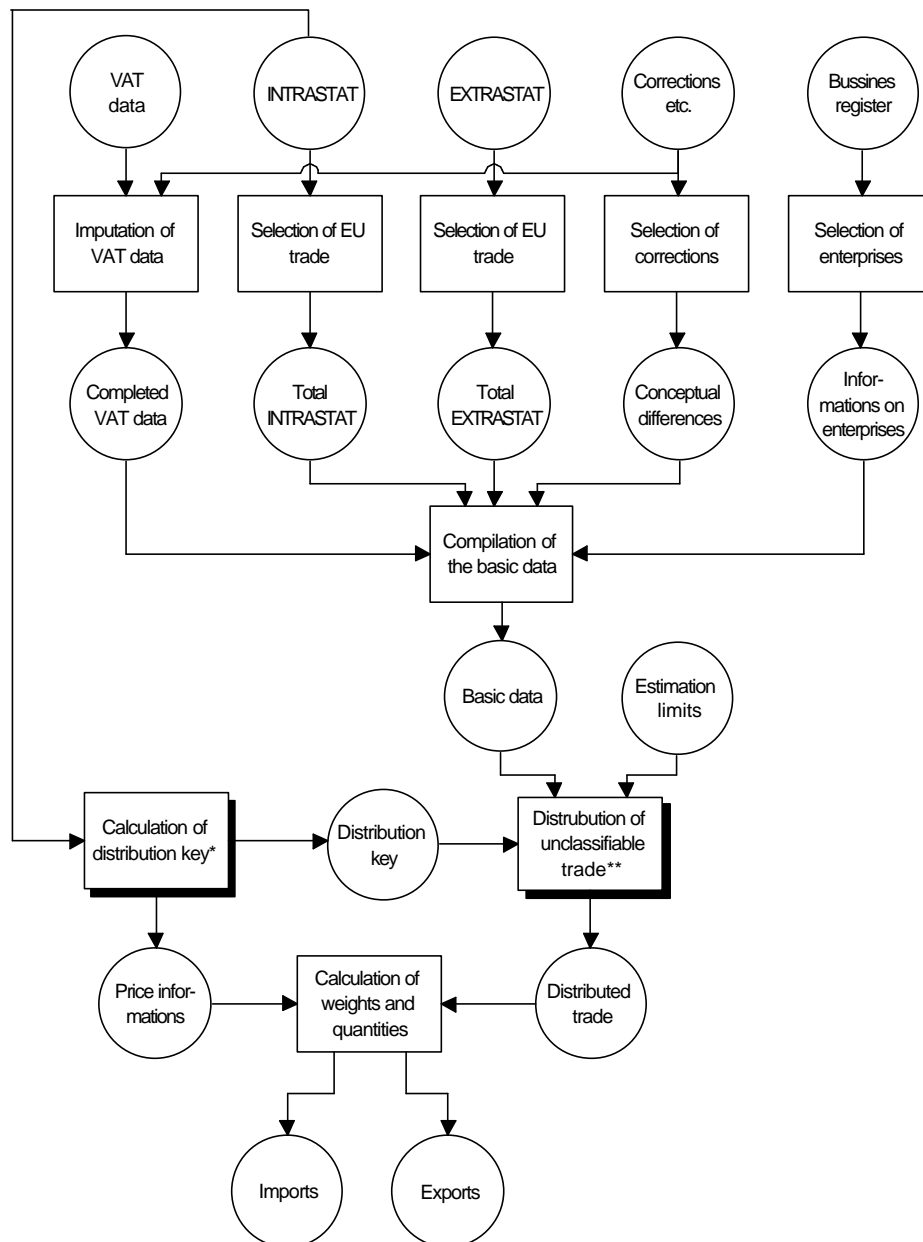
Thus, the non-reported trade can be distributed over countries and commodities for each enterprise in this group, in accordance with the distribution key representing the industry of the individual enterprise.

#### *Completion of data*

The enterprise specific contribution to the distribution of non-reported trade over countries and commodities is aggregated to one “enterprise“, representing the total amount of estimated trade. This exercise is primarily carried out in order to reduce the amount of data.

Based on the aggregated data, weights and supplementary units are subsequently calculated in accordance with the average unit values for each combination of country and commodity code (derived from the reported Intrastat data). This method ensures that there are no differences between the unit values concerning the distributed unclassifiable trade and the average reported unit values.

## Estimation and distribution of Unclassifiable Trade



\* The calculation of the distribution keys is described in Section 3.5

\*\* The distribution of the unclassifiable trade is described in Section 3.5

## 4. IMPUTATION PROJECTS IN EXTERNAL TRADE STATISTICS

No matter how well functioning, efficient and precise an imputation and estimation system is, there is always room for improvement. This is also the case for the system in the Danish External Trade Statistics. Below main areas of interest are shortly described, some of them including other administrative sources.

### 4.1 Use of data from Extrastat

Besides statistics on trade with the other Member States in EU, Statistics Denmark also produces and disseminate statistics on trade with non-EU countries (Extrastat). These data are collected by Customs, with the administrative purpose of controlling exports and collecting import duties. The data are transmitted to Statistics Denmark on a regularly basis as identifiable micro data for each company having

traded with non-EU countries. Statistics Denmark carry out a statistical error checking procedure, which results in regularly returns to the companies having reported the data.

Consequently, we are in the possession of time series of micro data, among other things showing the pattern of trade for each Danish company trading commodities across the Danish border, with both EU-countries and non-EU-countries. For companies historically having a similar trade pattern with EU and non-EU countries, possibly mostly regarding exports, the pattern from the non-EU trade data (Extrastat) could be used as an element in the imputation of commodity values (codes) in Intrastat. This case has its main relevance for companies lacking stability in their Intrastat declarations, because the Intrastat commodity distribution itself must be the preferred distributive variable in other cases.

#### **4.2 Use of data from the V.I.E.S. scheme**

An alternative administrative source that for many years has been considered interesting to use for imputation is data from the V.A.T. Information Exchange System (V.I.E.S. or Intra Community Sales List). In short all companies exporting to another EU-country have to report, for each recipient, the total value of export, the V.A.T. number of the recipient and the share of triangular trade.

As Intrastat the V.I.E.S. system was introduced in 1993 in connection with the internal market. The system was introduced in order to replace the border control undertaken before 1993. The purpose of the system was and is still to enable the Tax Authorities in the MS's to monitor if V.A.T is payed in the country of consumption as the V.A.T. law ascribe.

In theory the V.I.E.S data could be used as a similar measure of the total export as the V.A.T. information. Supplementary, if the data was exchanged between the MS's a supplementary measure could be calculated. One of the strengths with the V.I.E.S data compared with V.A.T. data, not taking into account possible differences in quality in the two systems, is the differentiation on recipients and hence country of destination. This information gives an indication of the distribution on partner countries, information which is not included in the V.A.T. data.

The use of V.I.E.S. data for imputation purposes has not yet been properly assessed but it is most certainly a priority for future studies with the aim to improve imputation.

#### **4.3 Imputation models and timeliness**

The increasing demand for earlier and more precise statistics entails a constant surveillance of the estimation model. If not, an earlier release of data will cause a more imprecise first estimate, and the element of unreliability will subsequently be reflected in the extent of the revisions.

The amount of reliable data is the main factor that must be dealt with, regarding timeliness in statistics. Two main aspects should be taken into account regarding this subject:

- Availability of alternative data sources (e.g. V.A.T. data) due to time lags. Due to timeliness in the data collection, the availability of alternative data sources will inevitable be reduced, if the first release of data is precipitated. Data must be collected and the quality should be appreciated before data can be used in the statistical models. Transmission routines between different authorities, is another time-consuming factor.
- Reduced data quality due to unfinished error-correction procedures. Timeliness in error-correction procedures will leave an increasingly number of possible errors unclarified.

As described, the administrative V.A.T. data is a major input in the models used to produce the external trade statistics. Regarding the V.A.T. data provided by the enterprises, the forms (paper or electronic) must be received at Customs within 25 days (monthly reporters) after the end of the reference period. Due to an increasing use of electronic V.A.T. forms, the coverage and the quality of the V.A.T. is assessed to be of a very high quality almost immediately after the reception. The 25 days deadline has

been in force since January 2004. This was a reduction of 15 days compared to earlier years, which has enabled the use of V.A.T. data in the Flash model.

After the reception of the V.A.T. reports, data are transmitted to Statistics Denmark after 30 days, obtaining approximately 98 per cent coverage of the monthly reports. In principle, data are also available after 25 days, which however will cause a reduction in the coverage to approximately 85 per cent. This reduction in the coverage must be dealt with, if the first release of data should be forwarded.

As the external trade statistics must be released within 40 days after the end of the reference month, this leaves only short time for error-correction procedures. A pre-control eliminates obvious errors, but no time is available for enterprise specific error-analyses. This implies a need for good imputation models regarding the imputation of missing V.A.T. reports (monthly, quarterly and half-yearly).

Similar, the deadline for the Intrastat forms (paper and/or electronic) is 10 days after the end of the reference period. The Intrastat reports are collected by Customs and forwarded to Statistics Denmark on a weekly basis not later than 14 days after the reception –14 days are used to register the detailed information from the paper forms in the computer system.

As data are transmitted each week, the amount of data will increase continuously. If all deadlines are kept, all data are, in principle, available for Statistics Denmark within one month from the end of the reference month. However, as described above, due to non-response and trade below the thresholds, only approximately 60 per cent of the data are available for the first release of statistics.

If the first release of the external trade statistics is forwarded, this coverage will be even further reduced, which inevitable will affect as well the aggregated as the detailed trade figures. Anyhow, the 60 per cent coverage indicates the urgent need for good imputation models.

In addition, as the timeliness of the Intrastat data implies, only very short time is available for error-correction procedures. Due to this, the quality of the data available for the first release data is rather low, compared to data that has passed through a full error-correction procedure (a full error-correction procedure requires approximately two weeks communication with the PSIs).

In order to reduce the non-response element, a reminder procedure is started, comprising all PSIs not fulfilling their obligation to Intrastat, i.e. in the case of total non-response. However, a standard reminder procedure lasts between two and six weeks, and therefore Intrastat for these PSIs will not be available until (at the earliest) the second release of the external trade statistics.

These circumstances is the primary reason, why the first release of data only comprises aggregated trade figures. This lack of detailed data underlines the importance of a qualified imputation model, and the development of such a model has high priority in the future.

Some initiatives towards this new model have already been taken. A Flash model based on V.A.T. data is planned to be implemented, and a new model regarding the imputation of the country-commodity matrixes is planned.

#### **4.4 Integration of the Flash and the Mester System.**

The Flash model used for the first figures for intra community trade has despite good results in the first four years some drawbacks. New information from alternative sources is not easily included and the estimates generated are on aggregated level.

The first problem relates to the more timely availability of information from the V.A.T. declaration. At the time of publishing the new trade figures the information on EU total buys and sales are now available contrary to before. This gives better information on the trade of enterprises since the early coverage is higher for the fiscal data on the V.A.T. declaration than in Intrastat.

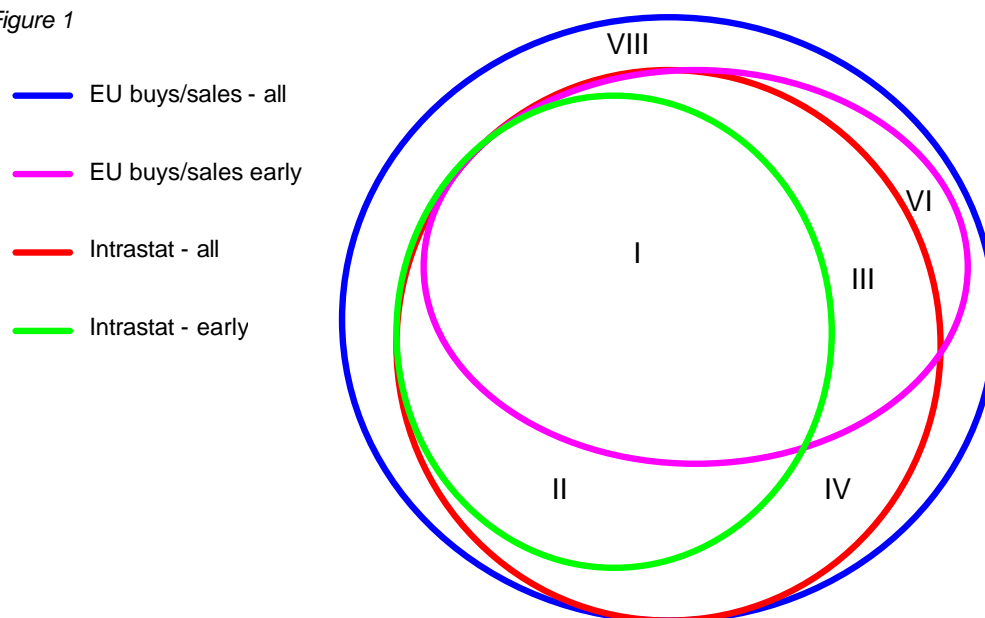
The latter problem is a question on the coherence between the Flash and the Mester models. Presently the Mester system uses all available enterprise specific information to distribute the imputations to commodity level. The difference in methods give rise to among other things major revisions on the detailed commodity level and problems in connection to the seasonally adjustment.

To improve the method and to reduce the problems between the models a more enterprise-oriented imputation is required in the Flash model.

A rough approach to achieve this is presented here. The model has not yet been completed and tested.

With the two sources Intrastat reports and V.A.T. data available eight subgroups can be identified for a specific reference month illustrated in the figure below. The blue circle in the figure represents the enterprises with intra-community trade. The red circle is the one that must report Intrastat. The pink and green are the enterprises that have reported EU buys/sales and Intrastat respectively at the time of the first publication.

Figure 1



## Enterprise groups

The eight groups are:

Group	In population	Reported Intrastat	Reported V.A.T.	Remarks
I	Yes	Yes	Yes	All information available
II	Yes	Yes	No	Only Intrastat available
III	Yes	No	Yes	Only V.A.T. available
IV	Yes	No	No	No information – complete non-response
V	No	Yes	Yes	In principle all information available but no trust to be placed in Intrastat
VI	No	No	Yes	Only V.A.T.
VII	No	Yes	No	No V.A.T. and no trust in Intrastat although available
VIII	No	No	No	No information at all

When these groups are identified the trade of five of them are determined immediately, namely the groups I, II, III, V and VI either by the reported Intrastat or the EU buys/sales. Before using EU buys/sales directly a correction should be made to take into account conceptual differences between the two concepts.

For the remaining three groups there are no information or only dubious information (group VII – Intrastat reported from enterprises not obliged to do so. The quality of data cannot be checked). The missing information for these groups will be determined by the trend of the known trade from the other five groups possible split by industry groups.

Using this method will ensure imputation on enterprise level and the method for spread on countries and commodities used in the Mester system can then directly be applied to the imputed data and ensure coherence between the Flash and Mester models.

-----