**Re-thinking Data Validation**

Anette Morgils Hertz, Statistics Denmark, ahz@dst.dk

Katja Overgaard, Statistics Denmark, kao@dst.dk

**Abstract**

*The 21st century has brought a number of challenges to data validation. At the office of External Economy at Statistics Denmark, we experience this in two ways: 1) Globalization leads to rapid changes in trade patterns, creating a strong need to ensure that we never miss major transactions by large multinationals. 2) Increasing demand to produce statistics of high quality with fewer resources available, forcing us to become better at prioritizing. In order to meet these challenges, External Economy has carried through a fundamental re-thinking of our validation routines. In this paper, we want to share the results.*

*The main point during the design of the new validation routines was to put a strong focus on the quality of the final/aggregate figures of each statistics. For International trade in goods for example, this means that focus is on the total figures reported by an enterprise, rather than the amount reported on each commodity code.*

*Secondly, when re-thinking our data validation it was important to ensure that we spent our time on the errors that matters the most. For example, a company can have such a big impact on the figures of one statistics, that it will be less urgent to investigate another enterprise’s outlier. To ensure that we are data driven when we determine which outliers to investigate we have created a key account system, where each of the probable errors is weighted across statistical domains, ensuring a clear priority of outliers to be investigated. The main finding has been that closely monitoring the validation and a key account approach makes the validation more efficient and more suitable for the challenges that we face.*

*In the paper, we introduce the changes to the validation process, as well as explain how we monitor the outcome of the various validation routines to ensure that we still perform optimally and reach our goal of focusing on the most important errors within a globalized economy with limited manpower available.*

**Keywords:** validation, efficient editing, quality, monitoring outcome, globalization

# 1. Introduction

When producing official statistics data validation and consistency checks are a big part of the production process before the final statistics can be disseminated. At the office of External Economy at Statistics Denmark we are responsible for disseminating the three statistics International trade in goods (ITGS), International trade in services (ITSS) and the Balance of payments (BoP). ITGS and ITSS are statistics with a lot of microdata, which is very time consuming to validate. In fact it is impossible to validate all. Hence, prioritizing which potential errors have to be validated each month is necessary.

A lot of the companies reporting their trade to ITGS and ITSS are producing globally. This means that the old trade patterns where a company produce in one country, and import and export goods and services to and from this country are changing. Hence, when validating data, specialised knowledge of global production setups is necessary.

In order to accommodate the need for specialised knowledge on globalised production and to ensure that our validation routines are always identifying the potential errors that should be validated each month, we have carried through a fundamental re-thinking of our validation routines at External Economy, where the main focus has been on the errors that matters the most on the final disseminated figures.

In the following sections the new validation routine will be introduced, as well as the system that has been developed to ensure our prioritising is always optimal given the resources available for data validation.

We also describe how our validation routines are monitored to ensure that the outcome of the validation routines is optimal, and that we reach our goal of focusing on the most important errors.

# 2. New validation routine

Prior to the re-design of the validation routines, there were separate validation routines for each of the three statistics disseminated by External Economy at Statistics Denmark. In fact we had 59 separate validation routines, distributed across ITGS, ITSS and BoP. This was the result of several decades’ development of validation routines. It was not the most effective way to work, as it often happened that the same companies were contacted by different clerical workers due to the same potential errors.

The new validation routine is simplified a lot, by dividing the current validation routines in company (PSI) specific checks and other checks. By implementing a key account approach to the PSI specific checks it is avoided that the same company is contacted by more than one clerical worker. The idea is that one person equals one company, such that a given company only have one contact at External Economy. In fact an IT system (called the key account system, described in section 4) has been developed to support this way of working.

In figure 1 the different validation routines are mapped. The PSI specific checks are divided in checks at the aggregated PSI-level, on the total import/export reported by a PSI, and at micro level on the units the PSIs report on (commodity codes, country codes, services etc.). The micro level PSI check is both a check for absolute errors and a check for probable errors. Absolute errors are for example a wrong format of data and probable errors are for example a wrong unit value of a certain good or service? For ITGS data the unit value check is called the SAND validation routine, and for ITSS the validation routine is actually one check combining the aggregated PSI level and the probable error check on microdata.

ITGS consists of Intrastat (trade with EU countries) and Extrastat (trade with non-EU countries). Hence in figure 1 it is shown that the PSI level checks are performed on both ITGS (Intra and Extra) and ITSS data.

The other checks we developed are not all relevant for both statistics. However, some are checks across the two domains, to ensure the quality of the BoP statistics, whereas other checks are only relevant for Intrastat or Extrastat. The other checks will not be described further in this paper, as most are not performed on a monthly basis, and hence the monitoring is simply a measure saying whether all potential errors over a certain value have been handled.

**Figure 1. Validation routines at External Economy**Source: External Economy, Statistics Denmark

In figure 2 it is shown how the validation work is distributed at External Economy. There is a SKV (statistic critical company) team, consisting of five people working part time on SKV validation, and a validation team consisting of four people doing fulltime validation.

**Figure *2*. Validation routines at External**Source: External Economy, Statistics Denmark

The main focus of the SKV team is to ensure that we do not miss major changes in the large multinational companies’ trade pattern. Hence our regular monthly validation checks are performed on the PSIs determined to be critical to the three statistics at External Economy. Also the most used codes for each SKV is checked monthly. If there are deviations from their regular trade patterns the SKVs are contacted. Annual and quarterly checks are performed across statistical domains on the SKV’s to ensure that the trade patterns reported to External Economy is the same as what is reported to other statistics.

The focus of the validation team is on all the other PSIs reporting to ITGS and ITSS. A key account system is designed to help priorities which PSI to control first. Since resources has been taken from the validation team in order to allow focus on the SKVs it has been necessary to increase the amount of automatic corrections, made on the PSIs that are not SKVs. This of course makes it necessary to also monitor the amount of automatic corrections made.

# 3. Monitoring the validation routines

As mentioned above the main focus when designing the new validation routine was on the errors that matter the most. This, we define as how big we believe the potential errors made on the final disseminated figures are.

The monitoring of the validation routines is twofold. There is the monitoring of the performance of each of the validation routines, and then there is the weighting between the validation routines in the key account system, to ensure we focus on the errors that potentially have the biggest impact on the final figures. Hence we use our manual resources optimally.

## *3.1. Monitoring the performance of the validation routines*

For each of our validation routines we have defined how we can evaluate the performance of the routine. Besides that we have a monitoring report that each month tells us how many automatic corrections have been made, how the SAND validation has performed, as well as what type and how many absolute errors there still are in data.

The validation routines are evaluated based on the hit rate calculated for each routine. The SAND validation routine has it incorporated, such that it can be calculated at all times, whereas we have not yet developed an automatic way to calculate a hit rate on the aggregated PSI level check and the ITSS combined check. Hence, a manual check of the two is performed once a year, to ensure they still perform optimally.

## *3.1.1. The hit rate explained*

The SAND validation routine is our probable error detection on ITGS PSI micro data level. It finds probable errors in the reported microdata, by looking at how suspicious a reported unit value is compared to the PSIs previously reported trade. The suspicion is multiplied by the potential impact the reported data has on the final figures. Around 1% of all data lines reported to ITGS is identified as a probable error via the SAND validation routine. The hit rate is calculated by taking the probable errors that turned out to actually be errors and divide by the total amount of identified probable errors. Our goal is to have a hit rate of 50%, such that the validation routine is correct at least half the time.

|  |  |  |
| --- | --- | --- |
|  | $$hit rate=\frac{actual errors}{identified probable errors}$$ | (1) |

The hit rate of the SAND validation routine is currently 51 %. This is slightly above our goal. Whereas the latest manual check of the hit rate on the aggregated PSI level check for ITGS is only 31% and 40% on the combined ITSS validation routine. This has led to a discussion of how the hit rate can be improved, by making the model better at identifying the correct outliers. As the aggregated PSI level checks are the validation routines that look at the PSIs that potentially impact the disseminated figures the most it is important that the validation routine is fine-tuned such that we focus on the figures that matter the most. At the moment we do not spend our time optimally. In fact we have asked our Methodological unit at Statistics Denmark to look at both validation routines in order to suggest improvements. In the meantime we are continuing to monitor the hit rate of the two routines, as the data collection on both is manual, and hence vulnerable to human error.

## *3.2. Monitoring report*

Each month a report is generated. It includes data on how many automatic corrections there have been compared to previous months. How many probable errors from the SAND validation routine there are still pending, how many have been accepted without corrections and data on how many have been corrected. The report also includes the amount and value of absolute errors, as data marked as absolute errors are not included in the disseminated figures.

At the end of each month we have a look at the reported data. We discuss whether everything seems to be as usual with the new data to be published. Part of this meeting is discussing, whether there has been anomalies in the monitoring report. If it is decided that something looks dubious, it is investigated immediately.

The monitoring report consists of several diagrams explained in the following. In figure 3 the amount of automatic corrections are depicted. We have two types of automatic corrections. One called 04 for short. It is automatic correction of probable errors from the SAND validation routine, as well as automatic corrections of absolute errors. The other form of automatic correction we call 06 for short. It is an automatic correction on either weight or supplementary unit of goods, where the reported unit value is in violation with the chosen commodity code or the customs procedure code is in violation with the chosen country code. In figure 4 the amount in DKK corresponding to the two different types of automatic correction are depicted. Nothing in the two figures looks concerning.

**Figure 3. Amount of automatic corrections** 

Source: Monitoring report, External Economy, Statistics Denmark

**Figure 4. Value in DKK of automatic corrections** 

Source: Monitoring report, External Economy, Statistics Denmark

In figure 5 the amount of probable errors from the SAND validation routine (marked as 21), the amount of probable errors that were accepted without correction (marked as 10) and the amount of probable errors that were corrected (marked as 12) is depicted. In 2019M01 there is a spike in the amount of probable errors that were accepted without corrections. This was investigated, and it turned out that by a manual mistake all records on a PSIs reported trade were accepted with code 10, where only one record was originally marked as a probable error. Unfortunately it was a report with around 1200 records, resulting in the big spike.

**Figure 5. Amount of potential errors from the SAND validation routine**

Source: Monitoring report, External Economy, Statistics Denmark

In figure 6, the amount in DKK is shown for the SAND validation routine. In 2018M06 there was a spike in the amount in DKK of the probable errors accepted without corrections. When investigated, it turned out that there were two very large records, marked as potential errors. In both instances the PSIs were contacted, and they confirmed that the reported data was correct.

**Figure 6. Value in DKK of potential errors from the SAND validation routine**

Source: Monitoring report, External Economy, Statistics Denmark

The amount of absolute errors and the corresponding value is captured in table 1. Since all absolute errors should be corrected automatically, we will investigate why there is one in 2019M05.

**Table 1. Overview of absolute error**

|  |  |  |
| --- | --- | --- |
| **Period** | **Value in DKK** | **Amount** |
| 2019M05 | 3.272.563 | 1 |

Source: Monitoring report, External Economy, Statistics Denmark

*3.3. Key account weighting*

The other aspect of our monitoring is the weighting between the validation routines in our key account system. The validation routines in the key account system are weighted such that the clerical workers always know which company to focus on first, as the companies are ranked according to how big an impact they have on the final figures.

In the key account system there are currently three different validation routines that are ranked automatically to ensure the focus of the validation team is put on the companies with the potential errors that have the biggest impact on the final figures.

The three validation routines are:

**CVRT** – aggregated PSI level check of ITGS reporters

**SAND** – A check for reasonable unit values on commodity codes (ITGS micro level PSI check)

**Txx** – Combined aggregated PSI level check and micro level check of ITSS reported data (xx is number from 00-06 indicating what type of errors there are on a report by a given PSI)

Each of the potential errors the routines return are ranked according to the potential error they respectively make on the final figures of ITGS and ITSS. In the key account system they are weighted together in such a way that the amount of errors we would have time to control each month, before the key account system was designed, is present. And by weighing the three validation routines together on PSI level it becomes clear which PSI should be contacted first each month.

The weighting system in the key account system is very flexible, which means it is easy to adjust if new validation routines are added, if the priority between the validation routines changes, if changes to the validation team is made and so on. This is done to ensure that we always use our resources optimally.

**4. Overview of the key account system**

In the following the key account system is described. As mentioned earlier it is developed in order to facilitate the new validation routines. The main idea behind the system is that one person handles all the PSI specific errors for one PSI. This gives the PSI one entry to External Economy and makes it possible for the person in charge of the PSI at External Economy, to look across different validation procedures. The key account system will be the main tool for the validation team in their everyday work with validation and the SKV team will also use it in their validation of the large PSIs critical to External Economy.

To make sure the system is user friendly and includes all necessary information, the users of the IT system have been involved in testing the program. The system will be continuously altered in connection with the optimization/simplification of validation routines. The user interface is similar to the current validation systems making the transformation from the old validation routines to the new less burdensome for the users of the system.

The IT architecture of the key account system is shown in figure 7. In addition to the validation systems that are related to ITGS, the key account system is developed to also be able to include validation systems that are related to ITSS. Thus, the clerical workers will have one joint system to evaluate both ITGS and ITSS. This is especially useful when considering validation routines that involve crosschecking of ITGS and ITSS. To incorporate information about the industry of the PSI in the key account system, the system will also get data from the business register.

**Figure 7. Architecture of the key account system**


Source: IT documentation of the key account system, External Economy, Statistics Denmark

The key account system has two main parts – an administration module and a part with the joint list of potential errors. The administration module is not expected to be used on a daily basis, where on the other hand, the joint list of potential errors is going to be the clerks’ everyday tool in their work.

In the administration module, as shown in the figure 8, one can specify which clerical workers should be assigned to which industries and thus the PSIs in that industry[[1]](#footnote-1). The system also allows individual PSI numbers to be allocated to clerks.

In the last part of the administration module one can specify which validation systems are to be included in the overall error score, which sources (Intra, Extra, etc.) that shall be drawn on and how much the individual validation systems shall weight relative to the total error score.

**Figure 8. Administration module of the key account system**


Source: The key account system, External Economy, Statistics Denmark

The other part of the system is a joint list of potential errors that draw on the company specific validation systems that are listed in the administration module. Figure 9 shows this joint list of potential errors in the left part of the screen. The list is ranked according to the error scores of current potential errors. In addition to the score, one can see which clerical worker is in charge of the potential errors, the company name, specific PSI identification number and which validation systems the record appears in.

On the right side of the screen the potential error which has been marked on the left side is shown. The right-side list shows all cases that the chosen company has in the various validation systems. Thus, it is possible to get an immediate overview of all errors that is detected for a specific company. The list is ranked by the score of the errors. When one clicks on a case on the company specific list, the domain specific validation IT system, where the case has arisen, opens i.e. UHTA (validation system for ITSS) or UHA (validation system for ITGS).

**Figure 9. List of cases in the key account system**

Source: The key account system, External Economy, Statistics Denmark

The new system, thus, gives our clerical workers a comprehensive overview of all potential errors they might ask the company about when contacting them. In that way, we avoid that several clerks contact the same company with questions regarding the same report, as well as we know what to ask about, as we have the complete overview of the company’s reported figures. It also ensures that the manual resources are spent optimally, as the clerical workers always know which PSI that potentially makes the biggest error on the final figures.

**5. How to change or update the list of cases**

The conditions for compilation of external trade statistic are changing over time. Both in terms of resources available for compilation and the fast changing trade patterns that arise from a more globalized world. This fosters a need for a validation system that is agile enough to follow along when these changes occur. The key account system’s administration module, described in the previous section, allows for some of these changes to be dealt with easy and fast such that we do not need to develop a new system or make major changes to existing ones.

The administration module in figure 8 shows the different options of how to change weights, assign specific PSIs to certain employees and divide enterprises among the staff based on their NACE code of economic activity[[2]](#footnote-2). In table 2 and 3 there is an example of what happens when the weights in the administration module is changed. A fictive example is constructed to overcome confidentiality concerns. Table 2 presents part of a list of potential errors where the total score is based on different weighted individual errors. The number in brackets following the type of error shows how many errors there are of the different types for a given PSI.

In table 2 the errors with the potentially biggest impact on the final figures would be those of company 1, hence the greatest effort to investigate and potentially solve these errors should be put here. Then follows company 2, potentially affecting the final figures the most, and so on. By solving one error for company 1 the total score for company 1 will decrease and the next time the key account system is opened company 1 might not be at the top of the list anymore. This makes the system very agile and able to continually adapt to the new information that occur.

**Table 2. Fictive list of potential errors before changes**

|  |  |  |
| --- | --- | --- |
| **Company** | **Type of error** | **Score** |
| Company 1 | CVRT, T04(3), T02(5), SAND(27) | 403 |
| Company 2 | CVRT, T04(3), T02(7) | 189 |
| Company 3 | T04(4) | 72 |
| Company 4 | SAND(5) | 40 |

Source: External Economy, Statistics Denmark

In figure 10 the weights have been changed in the administration module, resulting in a new list of potential errors presented in table 3.

**Figure 10. The administration module after changing the weights**



Source: The key account system, External Economy, Statistics Denmark

By changing the weights the list of potential errors could end up looking differently as shown in table 3. In this case, company 2 now potentially has the most important errors and we would solve these first. Next up is company 1 with the second most important errors to be solved. The scores for company 3 and 4 have changed, but the order of importance has not and they will still be looked at as third and fourth priority.

**Table 3. Fictive list of potential errors after changes in weights**

|  |  |  |
| --- | --- | --- |
| **Company** | **Type of error** | **Score** |
| Company 2 | CVRT, T04(3), T02(7) | 270 |
| Company 1 | CVRT, T04(3), T02(5), SAND(27) | 267 |
| Company 3 | T04(4) | 200 |
| Company 4 | SAND(5) | 13 |

Source: External Economy, Statistics Denmark

The above example shows how the key account system can be adapted to changes in validation procedures in the individual validation systems for ITGS and ITSS, which are providers of information to the key account system. This makes it agile to changes not only in underlying systems, but also if it is detected that one validation procedure needs more attention today, than it did when the key account system was first designed.

**7. Conclusion**

In this paper we introduced our re-designed validation routine. A re-design necessary due to the fact that we wanted to put a greater focus on validating the large multinationals, which globalised production setups require specialised knowledge. This change of focus required us to become better at prioritising the other potential errors we investigate each month – we want to spend our time on the potential errors that matter the most on the disseminated figures.

In order to ensure that our validation routines perform optimally and that we always focus our manpower on the potential errors with the greatest impact on our final figures we have designed at monitoring report, as well as an IT tool – the key account system, both introduced in the paper. The key account system allows a one clerical worker on company approach, ensuring focus on the potential errors on the final figures that matter the most. It is a very agile system, which makes it easy to adjust, if the Validation team changes, or if the priority between the different validation routine changes. The monitoring report helps us to be data driven, when evaluating each of our validation routines. With the key account system and the monitoring report we believe that we at External Economy are able to use our limited manpower at the most important errors.

1. Please note that some data is deleted for confidentiality reasons. [↑](#footnote-ref-1)
2. NACE is the ”Nomenclature generale des Activités économiques dans les Communautés Européennes” or in other words, the nomenclature of Economic Activities in the European Union. [↑](#footnote-ref-2)