Working Paper.

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Workshop on Statistical Data Editing
(Neuchâtel, Switzerland, 18-20 September 2018)

An application of selective and automatic editing to the Dutch International Trade in Goods Statistics

Prepared by Sander Scholtus, Bart Buelens, Jeffrey Hoogland, Jeroen Pannekoek and Rob Willems
(Statistics Netherlands)[1]

I. Introduction

1. The Dutch monthly International Trade in Goods Statistics (ITGS) describe the volumes of goods that are imported and exported by Dutch businesses, both in terms of quantity and value. In 2016, Statistics Netherlands (SN) started a re-design of the production process of these statistics. An important part of this re-design is focused on the editing process.

2. There are several incentives for SN to re-design the production of ITGS. In the context of editing, the main reasons are: (*i*) the availability of new data from administrative sources, which could be used to improve the efficiency of the editing process; and (*ii*) user demands to produce flexible output based on current events. The existing editing process was developed in the mid-1990s and was one of the first major production processes at SN to use macro-editing (Van de Pol, 1998). Our main aims for the new editing process are to improve the selection of influential errors by using administrative data as auxiliary information and to introduce automatic editing for all observations that are not edited manually.

3. The microdata for ITGS come from two main sources. For nearly all trade within the European Union (EU), data are collected through the so-called Intrastat survey. All Dutch businesses that trade goods within the EU beyond a minimal annual value (currently 1 million Euros for import and 1.2 million Euros for export) are required to report their transactions in the survey. In this paper we will focus on data that are reported on a monthly basis, which is the main component in terms of value. For other trade, including all trade with countries outside the EU, SN receives data directly from the customs authorities.

4. The microdata are reported as transaction data. The main output of ITGS consists of tables with total quantities and values of trade, aggregated across businesses and cross-classified by import/export, country of origin/destination, and type of goods. There are about 200 countries and (at the most detailed level) about 9500 type of goods codes according to Eurostat's so-called Combined Nomenclature. Hence, the potential number of cells in these tables is very large, although in practice many values will be zero. Note that zero values correspond to an absence of reported transactions, rather than zeros that are explicitly reported in the microdata.

5. In recent years, two administrative data sources have become available to SN which provide auxiliary information about businesses with international trade. Firstly, the total turnover from foreign trade is available in Value Added Tax (VAT) reports. This turnover value is subdivided by import/export and within EU/outside EU. However, it includes turnover from trade of services as well as goods. Secondly, for trade within the EU, businesses are required to file reports on 'intra-community supply of

[1] The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands. The authors would like to thank Jacco Daalmans for reviewing this paper.

goods and services' (Dutch abbreviation: ICP). This can be seen as a specification of VAT turnover from export by country (within the EU). In principle, turnover from import from EU countries is also available in this source, because it is reported as export by businesses in other EU countries. An advantage of ICP over VAT is that it contains separate values for goods and services. Neither administrative source is suitable as a direct replacement for the existing Intrastat survey or customs data since they lack the required level of detail. They do provide useful auxiliary information for editing. In the current process, these data are used on an ad hoc basis during manual editing. In the new process, SN aims to use them in a more systematic manner.

6. The remainder of this paper is organised as follows. In Section II, we provide a brief overview of the proposed new data editing process. In Section III, we describe the computation of anticipated values based on available auxiliary information, and the use of these values in a score function to identify influential suspicious observations. In Section IV, we propose a method for automatic editing and imputation of non-influential observations and non-responding units. To keep the exposition short, some complicating factors are ignored until Section V. Finally, some concluding remarks follow in Section VI.

II. Overview of the proposed data editing process

7. In broad terms, the proposed ITGS editing process would contain the following steps (in order):
   (a) Technical validity checks and adjustments
   (b) Correction of systematic errors
   (c) Computation of anticipated values
   (d) Automatic editing and imputation
   (e) Computation of selection scores for interactive editing
   (f) Interactive editing

Of these steps, only the last one involves direct human interaction.

8. Step (a) consists of basic technical checks on the input data before they are allowed to enter the production process. (For instance, do all type of goods codes refer to a valid item in the classification?) In step (b), rule-based editing is used to correct for specific errors that are known to occur and are relatively easy to recognise. (For instance, so-called 'thousand errors' where the reported values are 1000 times too large.) In the remainder of this paper, we will focus on steps (c) to (f). This part of the editing process is iterative: after a certain amount of interactive editing, an analyst can decide to re-run the automatic steps (c) to (e) again. In that case, the input for step (c) consists of an updated data set which still contains the improvements made by the editors, but not the automatic imputations from the previous iteration. Section III describes steps (c) and (e) in more detail. Step (d) will be the focus of Section IV. In the remainder of this paper we will focus on the editing process for trade values (i.e., amounts in Euros), not quantities.

III. Anticipated values and score functions

A. Anticipated values

9. As noted in the introduction, there are four main 'trade flows': import and export for countries within and outside the EU. These trade flows are treated separately in the production process and in the output. Throughout the remainder of this paper, we can therefore assume that we are working within one of these trade flows, without explicitly noting which one.

10. The most detailed aggregation level that we consider here is the total trade value in a particular month (t) of a particular business (b) for a particular combination of country (c) and type of goods code (g); denote this value as y_{tbcg}. Within a single business and month, these values can be aggregated to total values by country (y_{tbc+}), total values by type of goods (y_{tb+g}) and a global total value (y_{tb++}). Across businesses, monthly totals can be computed by combination of country and type of goods (y_{t+cg}) or even further by country (y_{t+c+}) or by type of goods (y_{t++g}). We also introduce the notation $y_{t\alpha}$ for a

generic value in month t, where the level of aggregation is not made explicit. Table 1 illustrates the different levels of aggregation for a single business in a single month.

Table 1. Overview of values for a single business (b) in a single month (t).

	type of goods					
country	1	2	…	$G-1$	G	total
1 2 … $C-1$ C	y_{tbcg}					y_{tbc+}
total	y_{tb+g}					y_{tb++}

11. For selective and automatic editing, we want to compare the observed values with anticipated values. Anticipated values are predictions for $y_{t\alpha}$, corresponding to expectations based on historic or auxiliary data. Different types of auxiliary information are available for the creation of anticipated values at the different aggregation levels in Table 1. For all cells, we can use historical edited data for the same business in previous months, i.e., y_{ubcg}, y_{ubc+}, y_{ub+g}, and y_{ub++} for months u in a certain reference period $T^*_{ref}(t)$. The reference period $T^*_{ref}(t)$ does not include month t itself, as this would lead to a circular argument. For future use, we also define an extended reference period $T_{ref}(t) = T^*_{ref}(t) \cup \{t\}$.

12. For trade flows within the EU, ICP data are available at the level of total values by country and the global total value, say z^{ICPG}_{tbc+} and z^{ICPG}_{tb++}. Here, the additional superscript 'G' refers to the goods component of ICP. The corresponding total ICP values, including trade of services, are denoted as z^{ICP}_{tbc+} and z^{ICP}_{tb++}. Finally, VAT data are available at the level of total values only, say z^{VAT}_{tb++}. As noted above, these values also include turnover from trade of services. If ICP data are also available for this business and trade flow, we can use these to estimate the share of VAT turnover that is due to trade of goods:

$$\tilde{z}^{VATG}_{tb++} = \begin{cases} \dfrac{z^{ICPG}_{tb++}}{z^{ICP}_{tb++}} z^{VAT}_{tb++} & \text{if } z^{ICP}_{tb++} \text{ is available and } z^{ICP}_{tb++} \neq 0, \\ z^{VAT}_{tb++} & \text{if } z^{ICP}_{tb++} \text{ is unavailable or } z^{ICP}_{tb++} = 0. \end{cases}$$

To create anticipated values for month t, we could use VAT and ICP values (if available) for month t as well as for previous months in $T^*_{ref}(t)$, that is to say, all months in the extended reference period $T_{ref}(t)$.

13. In practice, for any given business most of the values in Table 1, in particular in its interior, will equal zero. Moreover, this pattern of zeros need not be the same in different months, because a business might trade certain goods with certain countries on a less than monthly basis. These zero values are a complicating factor that is relevant for both selective and automatic editing. However, to keep the exposition simple, we will ignore them for most of this paper. We will return to this point in Section V.

14. Five simple methods to obtain an anticipated value for the total value y_{tb++} are as follows:
(1) Use the estimated VAT goods total for the same month ($\tilde{z}^{VATG}_{tb++}$).
(2) Compute the mean value of $\tilde{z}^{VATG}_{ub++}$ for months in the extended reference period ($u \in T_{ref}(t)$).
(3) Use the ICP goods total for the same month (z^{ICPG}_{tb++}).
(4) Compute the mean value of z^{ICPG}_{ub++} for months in the extended reference period ($u \in T_{ref}(t)$).
(5) Compute the mean value of y_{ub++} for months in the reference period ($u \in T^*_{ref}(t)$). Here, values based on automatic imputation are excluded from the computation.

Method (5) is, more or less, the method that is used to obtain anticipated values for macro-editing in the current process of ITGS at SN; cf. Van de Pol (1998).

15. For some businesses, the VAT and/or ICP values are systematically larger or smaller than the reported values in the Intrastat survey and customs data. To account for this, the first four methods in the above list can be extended by a simple ratio correction. Let $\hat{y}^{(m)}_{tb++}$ denote the anticipated value for y_{tb++}

according to method $m \in \{1,2,3,4\}$. As a robust correction factor for potential systematic deviations, we can take the median of the ratio of the edited values y_{ub++} and $\hat{y}_{ub++}^{(m)}$ in the reference period:

$$r_{tb++}^{(m)} = \underset{u \in T_{ref}^*(t)}{\text{median}} \left\{ y_{ub++} / \hat{y}_{ub++}^{(m)} \right\}.$$

The corrected anticipated value for y_{tb++} is then computed as $\hat{y}_{tb++}^{(m+5)} = r_{tb++}^{(m)} \hat{y}_{tb++}^{(m)}$. This yields four additional methods to obtain an anticipated value for y_{tb++}.

16. The above methods yield (at most) nine potential anticipated values for y_{tb++}. We want to choose the 'best' value from this set, i.e., the value that is likely to be closest to the unknown true value y_{tb++}^{true}. This can be quantified in different ways. After some experiments, we have adopted the following approach. We consider a potential anticipated value $\hat{y}_{tb++}^{(m)}$ to be 'reliable' if it deviates at most $q \times 100\%$ from the true value, i.e., if $\left| \hat{y}_{tb++}^{(m)} - y_{tb++}^{true} \right| < q \times y_{tb++}^{true}$. Here, q is a parameter between 0 and 1, for instance $q = 0.2$. Using the edited values y_{ub++} from the reference period $T_{ref}^*(t)$ as proxies for the associated true values y_{ub++}^{true}, the probability that $\hat{y}_{tb++}^{(m)}$ is a 'reliable' prediction can be estimated by

$$\hat{P}_{tb++}^{(m)} = \frac{1}{|T_{ref}^*(t)|} \sum\nolimits_{u \in T_{ref}^*(t)} I\left(\left| \hat{y}_{ub++}^{(m)} - y_{ub++} \right| < q \times y_{ub++} \right),$$

with $I(A) = 1$ if the event A is true and $I(A) = 0$ otherwise. As before, months in which y_{ub++} was obtained by automatic imputation are excluded from the computation of $\hat{P}_{tb++}^{(m)}$. The prediction $\hat{y}_{tb++}^{(m)}$ with the largest $\hat{P}_{tb++}^{(m)}$ is finally selected as the anticipated value for y_{tb++}, denoted as $\hat{y}_{tb++}$.

17. Figure 1 illustrates the computation of anticipated values for the total values of two different businesses. The values in this figure are fictitious but derived from actual data. Time series of the edited values, estimated VAT goods values and ICP goods values are plotted for 38 months from January 2014 onwards. The 'best' predicted values are plotted in blue. In the first example (left panel), the three auxiliary sources are usually in close agreement so the choice of method does not make much difference. In the second example (right panel), the edited values, VAT values and ICP values are significantly different throughout most of the period. Nevertheless, the predicted values are fairly reliable.

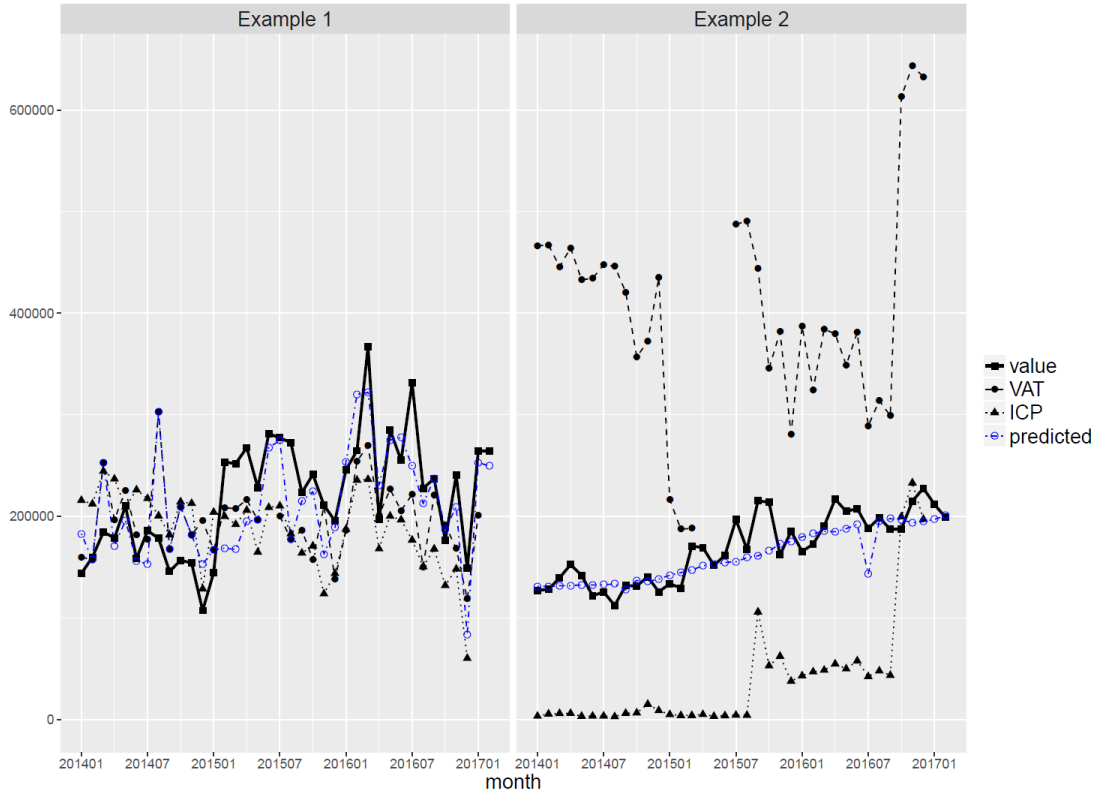

Figure 1. Two examples of the computation of anticipated values based on auxiliary information.

18. So far, we have focused on deriving anticipated values at the level of the total value of a business. The same approach can also be applied at the other levels of aggregation in Table 1, but in that case not all nine methods are available. For country totals within the EU (y_{tbc+}), methods 3, 4 and 5 and their ratio-corrected variants can be used, so it is still possible to choose the 'best' value. At other aggregation levels, only method 5 is available.

19. Since the anticipated values $\hat{y}_{tb++}$, $\hat{y}_{tbc+}$, $\hat{y}_{tb+g}$ and $\hat{y}_{tbcg}$ are often based on different sources of auxiliary information, these values usually do not satisfy the following basic accounting restrictions:

$$\begin{aligned} y_{tb++} &= \sum_c y_{tbc+} = \sum_g y_{tb+g}\,, \\ y_{tbc+} &= \sum_g y_{tbcg}\,, \\ y_{tb+g} &= \sum_c y_{tbcg}. \end{aligned} \tag{1}$$

This will become relevant when we consider their use in the context of imputation in Section IV.

B. Score functions for selective editing

20. To prioritise the most influential suspicious observations for interactive editing, a common approach is to use a score function (De Waal et al., 2011; Di Zio and Guarnera, 2014). A basic score function can be defined as the product of two components: a measure of the *risk* that an observed value is erroneous and a measure of the *influence* of that observation on target estimates. Observations with a high score are likely to contain influential errors and should therefore be prioritised for interactive editing.

21. To assess the risk of an observed value $y_{t\alpha}$, we start with the absolute deviation from its anticipated value: $\varepsilon_{t\alpha} = |y_{t\alpha} - \hat{y}_{t\alpha}|$. For a fair comparison of this deviation across different aggregates, it should be normalised by a measure of its expected size. That way, the accuracy of the anticipated value is taken into account, which may differ strongly between aggregates. As a risk measure, we propose:

$$R_{t\alpha}^{(1)} = \frac{\varepsilon_{t\alpha}}{M(\varepsilon_{t\alpha})} = \frac{|y_{t\alpha} - \hat{y}_{t\alpha}|}{M(|y_{t\alpha} - \hat{y}_{t\alpha}|)}, \tag{2}$$

where $M(\varepsilon_{t\alpha})$ is defined as the median of $\varepsilon_{u\alpha}$ for the edited values in the reference period ($u \in T_{ref}^{*}(t)$), with automatic imputations excluded.

22. As a measure of the (potential) influence of an observed value $y_{t\alpha}$, we take the maximum of this value and its associated anticipated value: $f_{t\alpha}^{(1)} = \max\{y_{t\alpha}, \hat{y}_{t\alpha}\}$. In this way, it is less likely that the influence of an observation is underestimated when the observed value is erroneously too small. The influence component is obtained by normalising $f_{t\alpha}^{(1)}$:

$$I_{t\alpha}^{(1)} = \frac{f_{t\alpha}^{(1)}}{F^{(1)}} = \frac{\max\{y_{t\alpha}, \hat{y}_{t\alpha}\}}{F^{(1)}}. \tag{3}$$

The normalising factor $F^{(1)}$ has to be determined separately for each aggregation level. We propose to define it such that $\sum_b I_{tb++}^{(1)} = 1$ and, within a given business, $\sum_c I_{tbc+}^{(1)} = \sum_g I_{tb+g}^{(1)} = \sum_c \sum_g I_{tbcg}^{(1)} = 1$.

23. By combining the risk (2) and the influence (3), we obtain the following score function:

$$S_{t\alpha}^{(1)} = I_{t\alpha}^{(1)} \times R_{t\alpha}^{(1)} = \frac{f_{t\alpha}^{(1)}}{F^{(1)}} \times \frac{\varepsilon_{t\alpha}}{M(\varepsilon_{t\alpha})}. \tag{4}$$

24. In the editing process as outlined in Section II, selection scores are computed *after* automatic editing and imputation. Consequently, the observed values $y_{t\alpha}$ used in (2)–(4) may have already been adjusted during automatic editing (to be discussed in Section IV). We would like to prioritise cases for

interactive editing where these (potentially adjusted) values are (still) suspicious according to score function (4), but we would also like to prioritise cases where automatic editing has introduced influential adjustments, even when the adjusted values themselves are no longer suspicious. This requires a second score function, which looks at the adjustments made during automatic editing.

25. As before, let $y_{t\alpha}$ denote a 'current' value, which may have been adjusted during automatic editing. Let $y_{t\alpha}^{base}$ denote the associated value at the start of step (c) in the first iteration of the editing process. To assess the effects of changes made during automatic editing, we define the following score:

$$S_{t\alpha}^{(2)} = \frac{|y_{t\alpha}^{base} - y_{t\alpha}|}{F^{(2)}}, \tag{5}$$

where the normalising factor $F^{(2)}$ is derived from the variable $f_{t\alpha}^{(2)} = \max\{y_{t\alpha}^{base}, y_{t\alpha}\}$ in the same way that $F^{(1)}$ is derived from $f_{t\alpha}^{(1)}$ in (3). We remark that $S_{t\alpha}^{(2)}$ in (5) can be seen as a score function of the same form as $S_{t\alpha}^{(1)}$, i.e., it can be written as a product of influence and risk components:

$$S_{t\alpha}^{(2)} = I_{t\alpha}^{(2)} \times R_{t\alpha}^{(2)}, \qquad I_{t\alpha}^{(2)} = \frac{f_{t\alpha}^{(2)}}{F^{(2)}}, \qquad R_{t\alpha}^{(2)} = \frac{|y_{t\alpha}^{base} - y_{t\alpha}|}{f_{t\alpha}^{(2)}}.$$

26. Finally, the two 'local' scores $S_{t\alpha}^{(1)}$ and $S_{t\alpha}^{(2)}$ have to be combined into a single 'global' score. Hedlin (2008) discusses a family of functions that can be used to obtain a global score. In the present case, it makes sense to take the maximum of the two score values, since we want to prioritise cases for interactive editing that have a high score on at least one of the two criteria:

$$S_{t\alpha} = \max\left\{S_{t\alpha}^{(1)}, S_{t\alpha}^{(2)}\right\}. \tag{6}$$

For values that have not been adjusted by automatic editing, $y_{t\alpha} = y_{t\alpha}^{base}$ and $S_{t\alpha} = \max\{S_{t\alpha}^{(1)}, 0\} = S_{t\alpha}^{(1)}$.

IV. Automatic editing and imputation

27. A new feature of the proposed editing process for ITGS is that data are edited automatically prior to selective editing. Given the dimensions of the data, analysts can check only a small subset of all possible aggregates. The score functions introduced in Section III are intended to guide this manual work so that it is focused on the most important and obvious anomalies in the data. By applying automatic editing to the rest of the data, we hope to improve the overall data quality. This way, the production of flexible output may become possible on short notice, with a minimum of additional interactive editing.

28. Automatic editing entails *error localisation* – i.e., identifying values that are deemed to be erroneous – and *imputation* of new values as a replacement for the identified errors. A common approach for automatic error localisation is based on the paradigm of Fellegi and Holt (1976): find the smallest (weighted) subset of values which can be imputed so that the resulting data are consistent with a set of restrictions for the data (*edit rules*). So-called *confidence weights* can be used to distinguish between values that are considered a priori more or less likely to be observed in error; a higher weight means that the value is less likely to be identified as erroneous.

29. For ITGS, it is natural to apply automatic editing separately to the data of each business for each month. These data have to satisfy a set of edit rules given by (1) and the restriction that all values at the lowest level of aggregation are non-negative: $y_{tbcg} \geq 0$. By themselves, these rules provide insufficient information to find errors: since businesses report only (positive) values for cells in the interior of Table 1 (y_{tbcg}) and the margins are derived during data processing, these edit rules are always satisfied by the observed data. To apply the Fellegi-Holt paradigm, we need additional edit rules.

30. Ideally, a value should be changed during automatic editing if it satisfies these two properties:
(1) The current value is suspicious.
(2) An accurate imputation is available as a replacement for the current value.

Here, we can define a value as being more 'suspicious' the more it deviates from its anticipated value, and we can define an 'accurate' imputation as one that is close to the (unknown) true value. With these definitions in place, it can be argued that a value $y_{t\alpha}$ is more likely to satisfy both properties as its risk score $R_{t\alpha}^{(1)}$ from (2) becomes larger. (Note: Here, $R_{t\alpha}^{(1)}$ is computed for the current value prior to automatic editing, whereas in Section III it was computed for the current value *after* automatic editing.)

31. Using the above argument, we proceed as follows. Choose two boundary values $0 < C_1 < C_2$ for the risk score and an interval $[w_{min}, w_{max}]$ for the confidence weights. We distinguish three cases:

(A) Current values $y_{t\alpha}$ with $R_{t\alpha}^{(1)} > C_2$ are always selected for imputation and receive the lowest possible confidence weight: $w_{t\alpha} = w_{min}$. We add the restriction that the risk score for the imputed value may be at most equal to C_2, i.e., the imputed value has to be less suspicious.

(B) Current values $y_{t\alpha}$ with $C_1 \leq R_{t\alpha}^{(1)} \leq C_2$ are not necessarily selected for imputation, but they may be selected because other values are imputed, in order to satisfy the restrictions in (1). These values receive a confidence weight in between w_{min} and w_{max} which decreases with $R_{t\alpha}^{(1)}$:

$$w_{t\alpha} = w_{max} - \frac{w_{max} - w_{min}}{C_2} R_{t\alpha}^{(1)}. \tag{7}$$

If such a value is selected for imputation, we add the restriction that the risk score for the imputed value may be at most equal to $R_{t\alpha}^{(1)}$, i.e., the risk may not increase due to imputation.

(C) Current values $y_{t\alpha}$ with $0 \leq R_{t\alpha}^{(1)} < C_1$ are not necessarily selected for imputation, but again may be selected in order to satisfy the restrictions in (1). These values also receive a confidence weight given by (7). In this case, *if* the value is selected for imputation, the risk score may be at most equal to C_1, i.e., the risk is allowed to increase due to imputation, but to a limited extent.

32. Each of the above restrictions on the risk score provides an interval of acceptable values for the imputed value, say $\tilde{y}_{t\alpha}$. More precisely, the restriction for case (A) that $R_{t\alpha}^{(1)} \leq C_2$ can be rewritten as:

$$\max\{0, \hat{y}_{t\alpha} - M(\varepsilon_{t\alpha})C_2\} \leq \tilde{y}_{t\alpha} \leq \hat{y}_{t\alpha} + M(\varepsilon_{t\alpha})C_2. \tag{8}$$

Note that the maximum in the lower bound arises from the restriction that $\tilde{y}_{t\alpha}$ cannot be negative. The restriction for case (B) that $R_{t\alpha}^{(1)}$ cannot increase beyond its current value is equivalent to:

$$\begin{array}{ll} \max\{0, 2\hat{y}_{t\alpha} - y_{t\alpha}\} \leq \tilde{y}_{t\alpha} \leq y_{t\alpha}, & \text{if } y_{t\alpha} \geq \hat{y}_{t\alpha}, \\ y_{t\alpha} \leq \tilde{y}_{t\alpha} \leq 2\hat{y}_{t\alpha} - y_{t\alpha}, & \text{if } y_{t\alpha} < \hat{y}_{t\alpha}. \end{array} \tag{9}$$

Finally, the restriction for case (C) that $R_{t\alpha}^{(1)} \leq C_1$ can be rewritten as:

$$\max\{0, \hat{y}_{t\alpha} - M(\varepsilon_{t\alpha})C_1\} \leq \tilde{y}_{t\alpha} \leq \hat{y}_{t\alpha} + M(\varepsilon_{t\alpha})C_1. \tag{10}$$

Thus, we have constructed a set of additional edit rules which can be used for error localisation.

33. As a small example, consider monthly trade values for a business, cross-classified by five countries and two types of goods. The left-hand table below shows observed values $y_{t\alpha}$ prior to editing; the right-hand table shows the anticipated values $\hat{y}_{t\alpha}$. Median prediction errors $M(\varepsilon_{t\alpha})$ are shown in brackets. It is seen that the total trade value is much lower than anticipated. The observed data satisfy all restrictions in (1), but the anticipated values do not, as explained above.

Observed data

	type of goods		
country	1	2	total
1	5	10	15
2	5	10	15
3	5	15	20
4	45	5	50
5	0	0	0
total	60	40	100

Anticipated values

	type of goods					
country	1		2		total	
1	5	*(2)*	5	*(2)*	10	*(2)*
2	10	*(3)*	10	*(4)*	20	*(3)*
3	20	*(10)*	20	*(10)*	30	*(4)*
4	5	*(3)*	35	*(5)*	40	*(4)*
5	10	*(2)*	100	*(5)*	110	*(4)*
total	55	*(3)*	200	*(5)*	250	*(5)*

34. From the information in the above tables, risk scores can be computed according to (2). These scores are shown in the left-hand table below. We have chosen $C_1 = 2$ and $C_2 = 10$ and colour-coded the cells according to the above three cases: red, orange and green scores represent cases (A), (B) and (C), respectively. The right-hand table shows the corresponding acceptable intervals according to (8), (9) or (10). By comparing the observed values to these intervals, it may be noted that case (C) values lie inside their acceptable intervals, case (B) values lie on the edge, and case (A) values lie outside their intervals.

Risk scores

	type of goods		
country	1	2	total
1	0.00	2.50	2.50
2	1.67	0.00	1.67
3	1.50	0.50	2.50
4	**13.33**	6.00	2.50
5	5.00	**20.00**	**27.50**
total	1.67	**32.00**	**30.00**

Acceptable intervals

	type of goods		
country	1	2	total
1	[1, 9]	[0, 10]	[5, 15]
2	[4, 16]	[2, 18]	[14, 26]
3	[0, 40]	[0, 40]	[20, 40]
4	**[0, 35]**	[5, 65]	[30, 50]
5	[0, 20]	**[50, 150]**	**[70, 150]**
total	[49, 61]	**[150, 250]**	**[200, 300]**

35. Let $\mathcal{A}_{tb}$ denote the index set of aggregate values $y_{t\alpha}$ for business b and month t. Define $d_{t\alpha} = 1$ if $\alpha \in \mathcal{A}_{tb}$ is to be imputed and $d_{t\alpha} = 0$ otherwise. The edit rules for the error localisation problem are given by: the consistency edits (1), non-negativity edits, and the acceptable intervals defined by (8), (9) and (10) for all $\alpha \in \mathcal{A}_{tb}$. According to the Fellegi-Holt paradigm, the following minimisation problem should be solved:

$\min_{d_{t\alpha}} \sum_{\alpha \in \mathcal{A}_{tb}} w_{t\alpha} d_{t\alpha}$, under the condition that all edit rules can be satisfied by imputing only values with $d_{t\alpha} = 1$.

However, this problem may not always have a feasible solution. In practice, the acceptable intervals for values at different levels of aggregation may sometimes contradict each other. (For instance, the lower bound on $\tilde{y}_{tbc+}$ might be larger than the sum of the upper bounds on $\tilde{y}_{tbcg}$, so that no values exist that satisfy all restrictions simultaneously.) In fact, these acceptable intervals are *soft* edit rules, which do not have to be satisfied in all cases. The Fellegi-Holt paradigm assumes that only *hard* edit rules occur.

36. Scholtus (2013) proposed an extension of the Fellegi-Holt paradigm which can accommodate soft edit rules. Here, each soft edit rule k is assigned a *failure weight* s_k. Define $\psi_k = 1$ if soft edit rule k is failed and $\psi_k = 0$ otherwise. The above minimisation problem is replaced by:

$\min_{d_{t\alpha}, \psi_k} \{\sum_{\alpha \in \mathcal{A}_{tb}} w_{t\alpha} d_{t\alpha} + \sum_k s_k \psi_k\}$, under the condition that all hard edit rules and soft edit rules with $\psi_k = 0$ can be satisfied by imputing only values with $d_{t\alpha} = 1$.

In the present context, we suggest to assign the same failure weight to each soft edit rule and to choose this weight much larger than w_{max}. Then, in practice, we will find the same solution as for the original Fellegi-Holt problem if that problem is feasible. If the original Fellegi-Holt problem is infeasible, this approach will yield a solution with as many values inside their acceptable intervals as possible.

37. Having obtained a solution to the error localisation problem, we need to find feasible imputations. (The Fellegi-Holt paradigm only guarantees that such imputations exist; it does not provide them.) Here, a two-step approach is often adopted (De Waal et al., 2011). We first create initial imputations by a simple method which does not take all restrictions into account. In the present case, it is natural to use the anticipated values as initial imputations. In the second step, the initial imputations are minimally adjusted (according to some criterion) to satisfy all restrictions. A common approach for this step uses a quadratic minimisation problem, which in this case takes the following form:

$\min_{\tilde{y}_{t\alpha}} \sum_{\alpha \in \mathcal{A}_{tb}} w_{t\alpha} (\tilde{y}_{t\alpha} - \hat{y}_{t\alpha})^2$, under the conditions that the adjusted values $\tilde{y}_{t\alpha}$ satisfy all hard edit rules and soft edit rules with $\psi_k = 0$, and that $\tilde{y}_{t\alpha} = y_{t\alpha}$ for all α with $d_{t\alpha} = 0$.

Note that only the imputed values may be adjusted in this step.

38. For the above example, we have set up the error localisation problem with confidence weights given by (7) with $w_{min} = 1$ and $w_{max} = 10$. In this example, it is possible to find an optimal solution where all values lie inside their acceptable intervals, i.e., no contradictory edit rules occur. In the left-hand table below, the observed values which were deemed erroneous have been deleted. Note that this

includes all case (A) values, as well as two case (B) values. In the right-hand table, the erroneous values have been replaced by their anticipated values. These initial imputations do not satisfy all restrictions.

Data after error localisation

	type of goods		
country	1	2	total
1	5	10	15
2	5	10	15
3	5	15	20
4	.	.	50
5	.	.	.
total	60	.	.

Data after initial imputation

	type of goods		
country	1	2	total
1	5	10	15
2	5	10	15
3	5	15	20
4	**5**	**35**	50
5	**10**	**100**	**110**
total	60	**200**	**250**

39. The left-hand table below shows the final imputed data, where the initial imputations have been adjusted by solving a quadratic minimisation problem. These values satisfy all edit rules. In particular, this means that no risk scores larger than $C_2 = 10$ occur and that all risk scores below $C_1 = 2$ have remained in that range. This is confirmed by the risk scores of the adjusted data in the right-hand table.

Final imputed data

	type of goods		
country	1	2	total
1	5	10	15
2	5	10	15
3	5	15	20
4	**25**	**25**	50
5	**20**	**115**	**135**
total	60	**175**	**235**

Updated risk scores

	type of goods		
country	1	2	total
1	0.00	2.50	2.50
2	1.67	0.00	1.67
3	1.50	0.50	2.50
4	6.67	2.00	2.50
5	5.00	3.00	6.25
total	1.67	5.00	3.00

40. In this example, the extended problem formulation with soft edit rules was not needed. However, suppose that the anticipated value for the top-left corner (country 1, type 1) was 100 instead of 5, still with a median prediction error of 2. The observed value 5 would then have a risk score of 47.50 and become a case (A) value with an acceptable interval [80, 120]. This interval is incompatible with the other anticipated values. By considering this interval as a soft restriction, the error localisation problem can still be solved. The resulting solution is the same as before, except that the value in the top-left corner retains its risk score of 47.50. This appears to be an acceptable solution in practice: the fact that the interval for this value is not in line with the other intervals might indicate a problem with this anticipated value, rather than with the observed value. If the value is influential, it may still be prioritised for follow-up during interactive editing.

41. We have created a prototype implementation of the above automatic editing strategy in R. This prototype uses the `editrules` package (De Jonge and Van der Loo, 2014) to solve the Fellegi-Holt based error localisation problem. As discussed in Scholtus (2015), the extended minimisation problem with soft edit rules can be rewritten in a form that can be handled by `editrules`. The quadratic minimisation problem to adjust the imputations is solved by the `rspa` package (Van der Loo, 2015). A variation on the above procedure has been developed for unit imputation of non-responding businesses.

V. Some complicating factors

42. An important complication that we have ignored until now in this paper is the occurrence of zero values in the data. In fact, many values $y_{t\alpha}$ are non-zero only in intermittent months. This has to be taken into account in the score functions and during automatic editing. Let $\pi^0_{t\alpha}$ denote the probability that a particular value $y^{true}_{t\alpha}$ is zero. This probability can be estimated by the relative frequency of zeros in the edited values $y_{u\alpha}$ for $u \in T^*_{ref}(t)$, say $\hat{\pi}^0_{t\alpha}$.

43. The expressions in Sections III and IV can be extended to account for the probability of a zero value. The absolute deviation $\varepsilon_{t\alpha}$ in (2) is replaced by $\varepsilon_{t\alpha} = \hat{\pi}^0_{t\alpha} y_{t\alpha} + (1 - \hat{\pi}^0_{t\alpha})|y_{t\alpha} - \hat{y}_{t\alpha}|$, which can be interpreted as an expected absolute deviation, given that $\hat{y}_{t\alpha}$ is the expected value when $y^{true}_{t\alpha} \neq 0$. For the resulting extended risk score, it is still possible to translate a restriction on the risk score into an interval of acceptable values for automatic editing, but the expressions become more complicated. For instance, for the restriction $R^{(1)}_{t\alpha} \leq C_2$, instead of (8) we find $L_{t\alpha} \leq \tilde{y}_{t\alpha} \leq U_{t\alpha}$, with

$$L_{t\alpha} = \begin{cases} \max\left\{0, \dfrac{(1 - \hat{\pi}^0_{t\alpha})\hat{y}_{t\alpha} - M(\varepsilon_{t\alpha})C_2}{1 - 2\hat{\pi}^0_{t\alpha}}\right\} & \text{if } \hat{\pi}^0_{t\alpha} < 1/2 \\ 0 & \text{if } \hat{\pi}^0_{t\alpha} \geq 1/2 \end{cases}$$

$$U_{t\alpha} = \begin{cases} M(\varepsilon_{t\alpha})C_2 + (1 - \hat{\pi}^0_{t\alpha})\hat{y}_{t\alpha} & \text{if } \hat{\pi}^0_{t\alpha} \leq 1/2 \text{ or } (\hat{\pi}^0_{t\alpha} > 1/2 \text{ and } \hat{\pi}^0_{t\alpha}\hat{y}_{t\alpha} \leq M(\varepsilon_{t\alpha})C_2) \\ \dfrac{M(\varepsilon_{t\alpha})C_2 - (1 - \hat{\pi}^0_{t\alpha})\hat{y}_{t\alpha}}{2\hat{\pi}^0_{t\alpha} - 1} & \text{if } \hat{\pi}^0_{t\alpha} > 1/2 \text{ and } \hat{\pi}^0_{t\alpha}\hat{y}_{t\alpha} > M(\varepsilon_{t\alpha})C_2. \end{cases}$$

Similar expressions can be derived to replace (9) and (10).

44. Other complicating factors are:

- For businesses that trade many different types of goods and/or with many different countries, the error localisation problem can become too large to be solved in an acceptable amount of time. This problem can be alleviated by considering separate error localisation problems for different clusters of types of goods. In a few exceptional cases, the problem has to be reduced further by focusing only on the largest values, so that the smallest values are not edited automatically.
- Sometimes the reliability of the anticipated values is difficult to assess, in particular for businesses that have non-zero values for a particular combination of country and type of goods only on a sporadic basis. The resulting score values for selective and automatic editing can be unreliable. We have introduced some modifications to the score functions to remedy this.

VI. Concluding remarks

45. In this paper we have described elements of a proposed new editing process for the Dutch ITGS. Anticipated values are obtained from several different, potentially conflicting sources of auxiliary information. These anticipated values are then used during selective editing, to prioritise influential suspicious values, and during automatic editing, both to define edit rules and as a starting point for imputation. An extension of the Fellegi-Holt paradigm that incorporates soft edit rules can be used for automatic error localisation. A prototype implementation was made using existing R packages. Currently, Statistics Netherlands is discussing the actual implementation of the new production process for ITGS.

References

de Jonge, E. and M. van der Loo (2014), *Error Localization as a Mixed Integer Problem with the Editrules Package*. Discussion Paper 2014-07, Statistics Netherlands, The Hague.

de Waal, T., J. Pannekoek, and S. Scholtus (2011), *Handbook of Statistical Data Editing and Imputation*. John Wiley & Sons, New York.

Di Zio, M. and U. Guarnera (2014), Theme: Selective Editing. In: *MEMOBUST Handbook on Methodology for Modern Business Statistics*, Eurostat, Luxembourg.

Fellegi, I. P. and D. Holt (1976), A Systematic Approach to Automatic Edit and Imputation. *Journal of the American Statistical Association* **71**, 17–35.

Hedlin, D. (2008), *Local and Global Score Functions in Selective Editing*. Working Paper No. 31, UN/ECE Work Session on Statistical Data Editing, Vienna.

Scholtus, S. (2013), Automatic Editing with Hard and Soft Edits. *Survey Methodology* **39**, 59–89.

Scholtus, S. (2015), *New Results on Automatic Editing using Hard and Soft Edit Rules*. Working Paper No. 35, UN/ECE Work Session on Statistical Data Editing, Budapest.

van de Pol, F. (1998), *Macro Editing in the Netherlands Foreign Trade Survey*. Research Paper, Statistics Netherlands, Voorburg.

van der Loo, M. (2015), *The rspa Package for Minimal Record Adjustment*. R Package, version 0.1.8.