UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS
Expert Meeting on Statistical Data Editing
3-7 October 2022, (virtual)

Towards a new integrated uniform production system for business statistics at Statistics Netherlands: automatic data editing with multiple data sources

Sander Scholtus, Wilco de Jong, Anita Vaasen-Otten and Frank Aelen, (Statistics Netherlands, The Netherlands)
s.scholtus@cbs.nl; w.dejong@cbs.nl; amvj.vaasen-otten@cbs.nl; f.aelen@cbs.nl;

I. Introduction

A. Towards our new production system

1. Statistics Netherlands is working on a new integrated uniform production system for business statistics. The main goals of this renewal program are more flexible output, a more agile and efficient production process, and facilitating more possibilities for innovations and further developments. These goals align with the need for more responsive output development and innovation, an efficient and future proof production process, ready for all kinds of data sources, and an increasing demand for new data, against limited internal resources. Our current processes are not agile enough to be able to respond timely to new developments and to new user needs. Agility is necessary, knowing that the complexity of economic phenomena increases, among others because of globalization and intertwining of industries. The current chain of economic statistics is designed in particular for current (mainly Eurostat) requirements. With the new integrated uniform production system the current high production workload will be reduced, so that staff can focus more on further innovation of processes and output.

2. Our innovation program started at the end of 2019 and is intended to run for five years. The program balances innovation, reuse of best practices and stepwise implementations. The innovations are based on the program principles in combination with applied methodology that is elaborated and tested in Proofs of Concept. The reuse of best practices is a co-creation with business and agile IT teams in order to realize standardized coherent models and common tools. The implementations take place in small steps instead of a big bang. The visible results encourage staff and the immediate feedback contributes to continuous further improvements.

3. Important aspects of the new production system are generalized modular building blocks with customizable settings (also reducing IT legacy burden) and the use of automatic data editing across multiple data sources (in addition to top-down manual data editing). For this modular setup, use is made of a suite of R packages for data editing, including `validate`, `errorlocate` and `simputation`. In the proposed new production system, automatic data editing will be performed not only within individual statistics, but also in an integrated manner for clusters of related statistics. This results in more consistent output across statistics and with National Accounts. It will also result in more efficiency because of the extra data available for business units and therefore the possibilities to identify additional rules for editing.

B. General principles of the new production system

4. In the development of the new production system, we started from the following eight principles. Fully implementing these principles in production for all business statistics is ongoing work that will take many years. In a companion paper by the same authors (Vaasen-Otten et al., 2022) the principles are described in more detail.

(1) We process our input automatically and immediately up to provisional output:
Input data from primary observations (questionnaires) and secondary sources (registers, registrations, web scraping and big data) are processed automatically, up to and including provisional output;

(2) We measure quality automatically and thus direct the manual work:
We continuously measure quality in an automated manner. We do this in the automated process steps as well as in the remaining manual work, so we can focus on actions that have the highest impact on the improvement of quality. For example, editing first takes place as much as possible automatically, where score functions are automatically calculated to measure the quality;

(3) We make our data consistent as early as possible:
Resolving errors as early as possible ensures a streamlined and efficient process. This prevents (multiple) processes further down the chain from being affected by these errors and prevents earlier process steps having to be performed again. Also, we prevent similar process steps from being performed multiple times by placing them as early as possible in the process. In the pursuit of consistency, the available data are therefore related as early as possible in the process;

(4) We share all our data, right from the start:
As soon as data arrives, it is immediately standardized, linked and made available to others. This concerns both primary data and secondary data. Availability of data is on a need-to-know basis;

(5) We centrally manage all our (population) frames, which are the basis of our statistics:
The (population) frames are centrally managed and made available for all units required to make our statistics. By central we mean in one place, but this may differ per frame. The (population) frames are accessible to everyone. The statistical units used for coordination are determined (limited set) and all incoming data is linked to those units;

(6) We have fully standardized our processes, methods, data and IT:
This principle consists of the following parts:
- All our processes are centrally described and coordinated in order to obtain optimal consistency;
- The metadata of all data is centrally described and managed;
- We work with coordinated statistical units;
- We log all our actions and this information is also available for further process optimization;

(7) Our processes, methods, data and IT are modular:
To be able to respond flexibly to new developments, our processes, methods, data and IT are modular, according to the Generic Statistical Business Process Model (mainly GSBPM phases 5-Process and 6-Analyse; https://statswiki.unece.org/display/GSBPM). An example is working with standard generic automatic editing modules, which are controlled by rules and can thus be quickly and flexibly adapted to specific situations;

(8) We resolve manual corrections the following iteration in the standard process:
Solutions for manual corrections that are necessary to improve quality, are incorporated in the next iteration in the standard process if possible, by adjusting sources, metadata, rules, parameters, questionnaires, and so on. This concerns, for example, frequently occurring manual corrections of data or inconsistencies between sources. This means we strive for a learning process.

C. Contents

5. The main focus of this paper is the automatic data editing with multiple data sources we intend to use in the new production system. Suppose that several micro datasets are available for business statistics, with some overlap between the units that are observed in each dataset. Furthermore, suppose that a number of *common variables* can be derived from these datasets, i.e., variables that can be compared directly across datasets because their definitions are sufficiently aligned so that, in the absence of measurement errors, the same unit should have approximately the same value in each dataset. The remaining variables cannot be compared directly across datasets, but many of them are in turn related to one or more common variables by edit rules. Rather than editing

each dataset in isolation, it seems advantageous to confront the microdata in these different sources at an early stage of the statistical process, as this may increase the overall quality of the edited data. Moreover, from the perspective of a statistical institute as a whole it may be more efficient and effective to identify inconsistencies between different datasets as soon as possible, and either resolve or explain them. Otherwise the underlying data problems may be resolved only partially and independently at several places in the office, which could create further inconsistencies that have to be addressed later, e.g., during the production of National Accounts.

6. The focus of this paper is specifically on ways to use the information on common variables during automatic editing. The use of common variables during top-down interactive editing is discussed in the companion paper Vaasen-Otten et al. (2022). The remainder of the present paper is organized as follows. Before addressing the problem of automatic editing across multiple data sources in Section III, we first give a brief review of existing methods for automatic editing of a single dataset in Section II. In Section IV, we discuss the use of a score function for identifying units with large inconsistencies between data sources, for which automatic editing may not be effective. Finally, the main conclusions are given in Section V.

II. Automatic editing: Review of existing methods

7. A widely used method for automatic editing consists of *deductive correction*. This approach is mostly based on correction rules (IF-THEN rules) that describe explicitly which adjustments have to be made to which variables under which conditions. As an example, consider the following rule that attempts to correct so-called thousand errors in current turnover values by looking at suspiciously large growth rates:

IF (turnover_current / turnover_previous > 300) THEN (turnover_current := turnover_current / 1000).

Deductive correction is most effective when it is used to correct errors with a systematic cause (De Waal et al., 2011). Its main advantage is that it allows a user to control precisely which adjustments are made to the data. If it is possible to translate the way subject-matter experts resolve data problems into IF-THEN rules, then deductive correction will yield edited data that resemble data that were edited by subject-matter experts. The main disadvantage of this approach is that a large and complicated set of correction rules may be required in practice to account for all possible situations that can occur. Such a large set of correction rules would be difficult to design and maintain (Chen et al., 2003).

8. For errors without a known cause, automatic error localization is often based on the paradigm of Fellegi and Holt (1976). This requires the specification of a set of *edit rules* which describe relations between variables in the dataset that should hold if the data are error-free. According to the Fellegi-Holt paradigm, the smallest possible subset of variables should be identified as erroneous for which it is possible to impute new values such that all edit rules are satisfied. More generally, so-called reliability weights can be assigned to distinguish between variables that likely contain more (lower weight) or fewer errors (higher weight). The error localization problem is then solved by minimizing the sum of reliability weights of the erroneous variables. Mathematically, this problem can be written as a mixed-integer programming problem, for which standard algorithms are available (De Waal et al., 2011; Van der Loo and De Jonge, 2018).

9. From a theoretical point of view, error localization based on the Fellegi-Holt paradigm is expected to work well when all errors affect one variable at a time and occur independently of each other (Liepins, 1980). With the application to multiple data sources in mind, this might limit its usefulness. Inconsistencies between data sources could indicate systematic issues that involve multiple variables. For instance, the reported values in one source might erroneously refer only to part of a unit and therefore be systematically smaller than values reported in other sources. This type of error is not likely to be identified correctly using the Fellegi-Holt paradigm.

10. To address the above limitation of the Fellegi-Holt paradigm, Scholtus (2016) and Daalmans and Scholtus (2018) proposed a generalization of this paradigm. Under this generalized paradigm, the (weighted) number of *edit operations* applied to a unit is minimized instead of the (weighted) number of erroneous variables. Here, an edit operation can be any operation that changes the values of one or more variables to new values that are a linear combination of the original values and zero, one or more free parameters. A set of admissible edit operations has to be selected beforehand. This set is application-specific and could be based on corrections that are commonly applied by subject-matter experts during interactive editing. Mathematically, the error localization problem with edit operations can still be written as a mixed-integer programming problem and solved in the same

way as before. The original Fellegi-Holt paradigm can be seen as a special case of this approach, for a particular choice of admissible edit operations (Scholtus, 2016).

11. From a simulation study, Daalmans and Scholtus (2018) concluded that edit operations can be useful as an extension to the error localization problem mainly if a user chooses edit operations that correspond to systematic errors that occur relatively often in the data. Like correction rules, edit operations aim to mimic how a subject-matter expert would handle particular data issues. With edit operations, the choice whether or not to apply a correction and the restriction that the corrected data have to be consistent with the edit rules are both handled by the error localization algorithm, which might make this approach more attractive than deductive correction, as the design and maintenance are less demanding. On the other hand, this approach does not provide users with the same level of control over which corrections are made during automatic editing, and users may perceive an error localization method as more of a 'black box' than using explicit correction rules.

12. In practice, deductive correction and error localization are often applied in tandem. A typical automatic editing process may contain the following steps in order (Pannekoek et al., 2013; GSDEM, 2019):

(1) deductive correction of systematic errors;
(2) error localization of other errors;
(3) imputation of missing values;
(4) adjustment of imputed values to satisfy all edit rules.

The first two steps have already been discussed. Imputation is a necessary step after error localization under the Fellegi-Holt paradigm, because this method only identifies certain variables as erroneous but it does not provide new values for these variables. In addition, some values in the data may have been missing from the start and also have to be imputed in step 3. As most common imputation methods do not take all edit rules into account (De Waal et al., 2011), some edit rules may still be failed by the data after step 3. In step 4, the imputed values are minimally adjusted according to some criterion (e.g., a weighted Euclidean distance) so that all edit rules become satisfied. Only values that have been imputed may be adjusted during this step; the error localization method from step 2 guarantees that it is possible to satisfy all edit rules in this way.

13. There exist various software packages that implement (part of) the above methodology for automatic editing. We will use a suite of R packages, developed at Statistics Netherlands, that provide an open source implementation of methods for all of the above process steps: `validate` for managing and evaluating edit rules throughout the process, `dcmodify` for deductive correction, `errorlocate` for error localization, `deductive` and `simputation` for imputation of missing values, and `rspa` for minimal adjustment according to a weighted Euclidean distance. We refer to Van der Loo and De Jonge (2018) for more information about these R packages.

III. Automatic editing of multiple data sources

A. Notation and problem statement

14. Suppose that there are P data sources that we want to edit in a consistent manner. We denote the vector of observed variables of unit i in source p as $\mathbf{x}_i^{(p)} = \left(x_{ij}^{(p)}\right)$, where $j \in \{1, \ldots, J^{(p)}\}$ indicates the variables in source $p \in \{1, \ldots, P\}$. For each data source, some internal edit rules may have been formulated. For notational simplicity, we assume that all edit rules are linear restrictions of the form

$$\sum_{j=1}^{J^{(p)}} a_{kj}^{(p)} x_{ij}^{(p)} \odot b_k^{(p)}, \tag{1}$$

where k is an index of edit rules, $a_{kj}^{(p)}$ and $b_k^{(p)}$ are given constants and the operator $\odot$ denotes either an equality ($=$) or an inequality sign ($\leq$ or $\geq$). More generally, conditional edit rules that consist of linear components (e.g., "IF $(x_1 > 0)$ THEN $(x_2 > 0)$") can also be handled during automatic error localization, by rewriting them in the form (1) using auxiliary binary variables; cf. Van der Loo and De Jonge (2018, Chapter 7).

15. Next, let $\mathbf{y}_i^{(p)} = \left(y_{il}^{(p)}\right)$ denote the common variables (as defined in Section I.C) that are observed for unit i in source p. In the simplest case, it holds that $y_{il}^{(p)} = x_{ij}^{(p)}$ for some $j = j(l)$. More generally, we assume that each common variable can be derived as a linear combination of observed variables in each source:

$$y_{il}^{(p)} = \sum_{j=1}^{J^{(p)}} c_{lj}^{(p)} x_{ij}^{(p)}, \tag{2}$$

for some known constants $c_{lj}^{(p)}$. Let L denote the total number of common variables between the available data sources. We denote the subset of all data sources in which common variable l is available for unit i as $B_{il} \subseteq \{1, \dots, P\}$. Note that B_{il} is unit-specific, since not every unit is necessarily observed in each data source, e.g., due to sampling and non-response. In practice, $L \ll \sum_{p=1}^{P} J^{(p)}$.

16. Given the assumption that the definitions of common variables are aligned between sources, the values $y_{il}^{(p_1)}$ and $y_{il}^{(p_2)}$ should be close together for each pair of sources (p_1, p_2) with $p_1 \in B_{il}$ and $p_2 \in B_{il}$. This motivates the formulation of additional edit rules across data sources of the form:

$$\left| y_{il}^{(p_1)} - y_{il}^{(p_2)} \right| \leq \epsilon_l \left| y_{il}^{(p_2)} \right|, \quad \text{for all } p_1 \neq p_2 \text{ with } p_1 \in B_{il} \text{ and } p_2 \in B_{il}. \tag{3}$$

Here, $0 \leq \epsilon_l < 1$ is a constant that defines the maximal allowed relative deviation between two values for common variable l. Below, we will use $\epsilon_l = 10\%$ for all common variables. Restrictions of the form (3) can be rewritten as conditional linear edit rules and therefore can be handled during automatic error localization.

17. In principle, the problem of making the available data of unit i consistent both within and across all data sources could now be solved as one large automatic editing problem: edit the data $\left(\mathbf{x}_i^{(1)}, \dots, \mathbf{x}_i^{(P)}, \mathbf{y}_i^{(1)}, \dots, \mathbf{y}_i^{(P)}\right)$ so that all restrictions (1), (2) and (3) are satisfied. In practice, this problem rapidly becomes too complicated and time-consuming as the number of data sources *P* and/or common variables *L* increases. We propose an alternative stepwise approach to solve this data editing problem, which should lead to a more manageable workload.

18. In order to set up this stepwise approach, it is useful to introduce a vector $\mathbf{z}_i = (z_{il})$ of 'true' values of the common variables for unit i. Given that the definitions of the common variables are aligned between sources, a unique 'true' value z_{il} conceptually exists for each unit, although it is not observed directly. (Of course, it may hold that $z_{il} = y_{il}^{(p)}$ for some p, but we do not know this a priori.) Instead of defining restrictions of the form (3), we can also restrict the relative deviations between observed values $y_{il}^{(p)}$ and their underlying 'true' value z_{il}:

$$\left| y_{il}^{(p)} - z_{il} \right| \leq \epsilon_l^* |z_{il}|, \quad \text{for all } p \text{ with } p \in B_{il}, \tag{4}$$

for some $0 \leq \epsilon_l^* < 1$. Suppose that we choose $\epsilon_l^* = \epsilon_l/(2 + \epsilon_l)$. Then it can be shown, using the triangle inequality, that any unit that satisfies all restrictions (4) also satisfies all restrictions (3):

$$\left| y_{il}^{(p_1)} - y_{il}^{(p_2)} \right| \leq \left| y_{il}^{(p_1)} - z_{il} \right| + \left| y_{il}^{(p_2)} - z_{il} \right| \leq 2\epsilon_l^* |z_{il}| \leq \frac{2\epsilon_l^*}{1-\epsilon_l^*} \left| y_{il}^{(p_2)} \right| = \epsilon_l \left| y_{il}^{(p_2)} \right|.$$

For the third inequality, it was used that (4) implies that $|z_{il}| \leq \left| y_{il}^{(p_2)} \right| + \left| z_{il} - y_{il}^{(p_2)} \right| \leq \left| y_{il}^{(p_2)} \right| + \epsilon_l^* |z_{il}|$. In practice, when $\epsilon_l \ll 1$ the simple choice $\epsilon_l^* = \epsilon_l/2$ works well as an approximation of $\epsilon_l^* = \epsilon_l/(2 + \epsilon_l)$.

19. A further advantage of introducing the 'true' values z_{il} is that this makes it easy to define additional edit rules of the form (1) that describe relations between common variables:

$$\sum_{l=1}^{L} a_{kl} z_{il} \odot b_k. \tag{5}$$

A simple, but useful example is a non-negativity restriction: $z_{il} \geq 0$. Note that for any common variable to which a non-negativity restriction applies, edit rule (4) simplifies to $(1 - \epsilon_l^*) z_{il} \leq y_{il}^{(p)} \leq (1 + \epsilon_l^*) z_{il}$.

B. Summary of proposed approach

20. The goal of consistent automatic editing within and across data sources can now be stated succinctly as: adjust the data $\left(\mathbf{x}_i^{(1)}, \dots, \mathbf{x}_i^{(P)}, \mathbf{y}_i^{(1)}, \dots, \mathbf{y}_i^{(P)}, \mathbf{z}_i\right)$ as necessary so that all restrictions (1), (2), (4) and (5) are satisfied. We propose to solve this automatic editing problem in three steps:

(1) Automatic editing of common variables across data sources
Errors are identified in $\left(\mathbf{y}_i^{(1)}, \ldots, \mathbf{y}_i^{(P)}, \mathbf{z}_i\right)$ in such a way that all restrictions (4) and (5) can be satisfied.

(2) Deriving additional edit rules on common variables
First the 'true' values in $\mathbf{z}_i$ are imputed, taking into account all restrictions (4) and (5). Next, given these imputed 'true' values, additional edit rules are derived for the common variables $\left(\mathbf{y}_i^{(1)}, \ldots, \mathbf{y}_i^{(P)}\right)$ in the form of lower and upper bounds that follow from (4).

(3) Automatic editing within each individual data source
The final step is carried out for each data source separately. For each $p \in \{1, \ldots, P\}$, errors are identified and new values are imputed in the data $\left(\mathbf{x}_i^{(p)}, \mathbf{y}_i^{(p)}\right)$ in such a way that all restrictions (1) and (2) are satisfied as well as all additional edit rules that were derived for $\mathbf{y}_i^{(p)}$ in step 2.

In the following subsections, each of these steps will be discussed in more detail.

C. Step 1: Automatic editing of common variables across data sources

21. The goal of this step is to identify errors in the common variables $\left(\mathbf{y}_i^{(1)}, \ldots, \mathbf{y}_i^{(P)}, \mathbf{z}_i\right)$ using all restrictions of the form (4) and (5). As part of this step we can apply both deductive correction and automatic error localization (see Section II). If it is reasonable to assume that the observed data in different sources were created by independent processes, then it follows that measurement errors in observed common values $y_{il}^{(p)}$ with the same underlying 'true' value z_{il} occur independently of each other. As this corresponds to one of the assumptions of the original Fellegi-Holt paradigm, error localization under this paradigm may then be an effective approach here. In addition, deductive correction and/or generalized edit operations could be useful to handle systematic errors that affect multiple common variables within the same data source, such as thousand errors.

22. To apply error localization in this step, reliability weights should be assigned to all common variables in $\mathbf{y}_i^{(1)}, \ldots, \mathbf{y}_i^{(P)}$. (Note: It is not necessary to assign weights to the 'true' values in $\mathbf{z}_i$ in this step, as these values are always missing.) These weights could be based on subject-matter knowledge, by comparing the reliability of each common variable across the data sources in which it occurs. If sufficient information is available, these weights could be differentiated by making them dependent on known characteristics of a unit such as its type of economic activity, size class and legal structure. A further refinement could be to derive reliability weights automatically for each unit using local scores (cf. Vaasen-Otten et al., 2022) as input.

23. The output of this step is a dataset for unit i in which it is possible to impute the missing values in $\left(\mathbf{y}_i^{(1)}, \ldots, \mathbf{y}_i^{(P)}, \mathbf{z}_i\right)$ in such a way that all restrictions (4) and (5) are satisfied. To achieve this, if necessary, some originally observed values in $\mathbf{y}_i^{(1)}, \ldots, \mathbf{y}_i^{(P)}$ may have been changed or set to missing. In particular, it is guaranteed that, after appropriate imputation of missing values, no deviations larger than $\epsilon_l |y_{il}^{(p_2)}|$ will occur between any pair of common variables $\left(y_{il}^{(p_1)}, y_{il}^{(p_2)}\right)$ with $p_1 \in B_{il}$ and $p_2 \in B_{il}$. The complexity of the error localization problem that has to be solved in this step remains limited as the number of sources P increases, because from each data source only the common variables $\mathbf{y}_i^{(p)}$ are taken into account, rather than all observed variables $\mathbf{x}_i^{(p)}$.

24. A small example is shown in Table 1. This example involves $P = 5$ data sources and $L = 5$ common variables for an enterprise in the manufacturing sector. This is only a subset of all data sources and common variables that were investigated during our Proof of Concept (see Section V below), to keep the example concise. The enterprise has been observed in: (1) a survey on Structural Business Statistics (SBS); (2) a survey on industrial production (ProdCom); (3) an administrative dataset on finances of small enterprise groups (Dutch abbreviation: SFKO); (4) an administrative dataset on Short-Term Statistics (STS-admin); (5) a survey on Short-Term Statistics (STS-survey). All data in Table 1 have been aligned to refer to annual values at the enterprise level.

Table 1: Example of automatic editing across data sources (steps 1 and 2). A dot (.) denotes a missing value.

variable	description	weight	original value	after step 1	after step 2
$y_{i1}^{(1)}$	Total turnover, definition 1 (SBS)	5	16 091	16 091	16 091
$y_{i1}^{(2)}$	Total turnover, definition 1 (ProdCom)	3	2 504	.	.
z_{i1}	Total turnover, definition 1 ('true' value)	-	.	.	**16 091**
$y_{i2}^{(1)}$	Total turnover, definition 2 (SBS)	5	16 091	.	.
$y_{i2}^{(3)}$	Total turnover, definition 2 (SFKO)	3	18 496	18 496	18 496
$y_{i2}^{(4)}$	Total turnover, definition 2 (STS-admin)	7	19 386	19 386	19 386
$y_{i2}^{(5)}$	Total turnover, definition 2 (STS-survey)	3	16 610	.	.
z_{i2}	Total turnover, definition 2 ('true' value)	-	.	.	**19 386**
$y_{i3}^{(1)}$	Total industrial production (SBS)	4	16 091	16 091	16 091
$y_{i3}^{(2)}$	Total industrial production (ProdCom)	6	15 582	15 582	15 582
$y_{i3}^{(5)}$	Total industrial production (STS-survey)	3	16 610	16 610	16 610
z_{i3}	Total industrial production ('true' value)	-	.	.	**16 091**
$y_{i4}^{(1)}$	Domestic industrial production (SBS)	4	16 091	16 091	16 091
$y_{i4}^{(2)}$	Domestic industrial production (ProdCom)	8	15 582	15 582	15 582
z_{i4}	Domestic industrial production ('true' value)	-	.	.	**15 582**
$y_{i5}^{(1)}$	Domestic industrial production of goods (SBS)	4	16 091	.	.
$y_{i5}^{(2)}$	Domestic industrial production of goods (ProdCom)	8	2 504	2 504	2 504
z_{i5}	Domestic industrial production of goods ('true' value)	-	.	.	**2 504**

25. The five common variables in this example are related to industrial production and total turnover (according to two different definitions). The following restrictions of the form (5) have been defined:

- $z_{i1} \leq z_{i2}$ (turnover according to second definition is never smaller than according to first definition);
- $z_{i3} \leq z_{i1}$ (turnover is never smaller than total industrial production);
- $z_{i4} \leq z_{i3}$ (total industrial production is never smaller than domestic industrial production);
- $z_{i5} \leq z_{i4}$ (domestic industrial production is never smaller than domestic industrial production of goods);
- $z_{i5} \geq 0$ (domestic industrial production of goods is non-negative).

Restrictions of the form (4) have been defined with $\epsilon_l^* = 5\%$, corresponding to $\epsilon_l \approx 10\%$ in (3).

26. The originally observed values of all common variables are shown in the fourth column of Table 1. It is seen that some inconsistencies with respect to (3) occur for this unit; for instance, the total turnover according to ProdCom is much smaller than according to the other sources. We have applied error localization based on the Fellegi-Holt paradigm to address these inconsistencies. The reliability weights are shown in the third column of Table 1. For SBS these weights were readily available from the automatic editing process that is currently used in production at Statistics Netherlands. For the other sources, reliability weights were set by subject-matter experts by comparing the expected quality of each common variable to the corresponding SBS variable. The output of error localization in this step is shown in the fifth column of Table 1.

D. Step 2: Deriving additional edit rules on common variables

27. In this step, we first impute possible values for the 'true' values z_{il} of the common variables. To do this, we begin by deriving lower and upper bounds on each z_{il}, given that (4) and (5) have to be satisfied by these values along with all observed values $y_{il}^{(p)}$ that were not identified as erroneous in step 1. Appropriate bounds can be derived by solving two linear optimization problems for each z_{il}, in which the value of z_{il} is minimized (maximized) given the restrictions (4) and (5). In R, this can be done automatically using the function `detect_boundary_num` from the `validatetools` package.

Table 2: Example of automatic editing across data sources (continued): feasible intervals for z_{il}.

variable	description	lower bound	upper bound
z_{i1}	Total turnover, definition 1 ('true' value)	15 819.05	16 937.89
z_{i2}	Total turnover, definition 2 ('true' value)	18 462.86	19 469.47
z_{i3}	Total industrial production ('true' value)	15 819.05	16 402.11
z_{i4}	Domestic industrial production ('true' value)	15 324.76	16 402.11
z_{i5}	Domestic industrial production of goods ('true' value)	2 384.76	2 635.79

28. For the example from Table 1, Table 2 shows the resulting lower and upper bounds on $z_{i1}, \ldots, z_{i5}$. As long as the imputed value for z_{il} conforms to these bounds, none of the non-missing values $y_{il}^{(p)}$ in the fifth column of Table 1 deviates by more than $\epsilon_l^* = 5\%$ from it. The restrictions (5) have also been taken into account.

29. Next, imputed values for z_{il} are derived that conform to the specified lower and upper bounds. The last column of Table 1 shows possible imputed values for the above example. In this small example, these imputations were derived by a simple ad hoc procedure: for each l we imputed z_{il} by the observed value $y_{il}^{(p)}$ of the variable with the largest reliability weight that remains non-missing after step 1 *and* does not violate the bounds. In general, this ad hoc procedure does not always yield appropriate imputations and a more advanced method is needed. In general, imputations could be derived from a statistical model and adjusted to satisfy the bounds.

30. Given the imputed values for z_{il}, we then derive lower and upper bounds on the common variables $y_{il}^{(p)}$ from the restrictions (4). These bounds are used to define additional edit rules for these variables of the form

$$L(z_{il}) \leq y_{il}^{(p)} \leq U(z_{il}), \quad \text{for all } p \text{ with } p \in B_{il} \text{ and } |B_{il}| \geq 2. \tag{6}$$

In the next step, these additional edit rules are used as restrictions to ensure that no new inconsistencies can be introduced between data sources when editing the individual sources. In the special case that $z_{il} \geq 0$, edit rule (6) has a simple form with $L(z_{il}) = (1 - \epsilon_l^*)z_{il}$ and $U(z_{il}) = (1 + \epsilon_l^*)z_{il}$. More generally, $L(z_{il})$ and $U(z_{il})$ can be derived as before, using the function `detect_boundary_num` from `validatetools`. The condition that $|B_{il}| \geq 2$ is added in (6) because restrictions of this form are considered informative only for those units that have observations on a common variable in at least two sources. Otherwise, no confrontation between observed values from different sources has actually taken place for variable l during step 1, and by using (6) we would simply restrict the value of $y_{il}^{(p)}$ to remain close to its original value in the only available data source.

31. For the above example, since all common variables are restricted to be non-negative and $\epsilon_l^* = 5\%$, all restrictions (6) take the simple form $0.95z_{il} \leq y_{il}^{(p)} \leq 1.05z_{il}$. For instance, for $y_{i2}^{(1)}$, turnover according to the second definition as observed in the SBS survey, we have the additional edit rule: $18\,416.7 \leq y_{i2}^{(1)} \leq 20\,355.3$.

E. Step 3: Automatic editing within each individual data source

32. In the final step, automatic editing is applied to each data source separately. For each p, the data $\left(\mathbf{x}_i^{(p)}, \mathbf{y}_i^{(p)}\right)$ are made consistent with the edit rules (1), (2) and (6). It is possible that new errors are found in $\mathbf{y}_i^{(p)}$ during this step, in addition to the errors that were previously found during step 1. The restrictions (6) are included to ensure that such errors are corrected without introducing inconsistencies with respect to the other data sources.

33. During this step, both deductive correction and error localization can be used as discussed in Section II. For error localization, the reliability weights of the variables in $\mathbf{x}_i^{(p)}$ can be chosen just as they would be for automatic editing within source p. For the common variables in $\mathbf{y}_i^{(p)}$, we propose to set the reliability weight to a very small value (e.g., 0.01), so that these variables may always be adjusted if this is necessary to satisfy the internal edit rules (1) and (2). The computational complexity of error localization in this step should be similar to that of error localization for one data source on its own.

34. Continuing the example from Tables 1 and 2, Table 3 contains part of the results of step 3 for the first data source, the SBS survey. It should be noted that in reality the SBS survey contains over 100 variables. In Table 3, we only show seven SBS variables $x_{ij}^{(1)}$ that are directly related to the five common variables $y_{il}^{(1)}$ used in the example so far. The relations (2) between these SBS variables and the common variables are as follows:

- $y_{i1}^{(1)} = x_{i1}^{(1)} - x_{i2}^{(1)} - x_{i3}^{(1)}$;
- $y_{i2}^{(1)} = x_{i1}^{(1)}$;
- $y_{i3}^{(1)} = x_{i4}^{(1)} + x_{i6}^{(1)} - x_{i2}^{(1)} - x_{i3}^{(1)}$;
- $y_{i4}^{(1)} = x_{i5}^{(1)} + x_{i7}^{(1)} - x_{i2}^{(1)} - x_{i3}^{(1)}$;
- $y_{i5}^{(1)} = x_{i5}^{(1)} - x_{i2}^{(1)} - x_{i3}^{(1)}$.

Table 3: Example of automatic editing across data sources (continued; step 3 for SBS).

variable	description	weight	original value	after step 2	after step 3
$y_{i1}^{(1)}$	Total turnover, definition 1	0.01	16 091	16 091	16 091
$y_{i2}^{(1)}$	Total turnover, definition 2	0.01	16 091	.	**18 417**
$y_{i3}^{(1)}$	Total industrial production	0.01	16 091	16 091	16 091
$y_{i4}^{(1)}$	Domestic industrial production	0.01	16 091	16 091	16 091
$y_{i5}^{(1)}$	Domestic industrial production of goods	0.01	16 091	.	**2 629**
$x_{i1}^{(1)}$	Total turnover, SBS definition	5	16 091	16 091	**18 417**
$x_{i2}^{(1)}$	Excises paid	3	0	0	**2 326**
$x_{i3}^{(1)}$	On-charged freight costs	4	0	0	0
$x_{i4}^{(1)}$	Total industrial production of goods, SBS definition	1	16 091	16 091	**4 955**
$x_{i5}^{(1)}$	Domestic industrial production of goods, SBS definition	1	16 091	16 091	**4 955**
$x_{i6}^{(1)}$	Total industrial production of services, SBS definition	1	0	0	**13 462**
$x_{i7}^{(1)}$	Domestic industrial production of services, SBS definition	2	0	0	**13 462**

35. It is seen that one of the corrections made during automatic editing in this example is that part of industrial production of goods as reported in the SBS survey is moved to industrial production of services. Looking back at Table 1, it is seen that this correction removes an inconsistency between SBS and ProdCom. It should be noted that a special edit operation (cf. Section II) was included here during error localization of SBS data that moves an arbitrary amount from industrial production of goods to services or vice versa. The other data sources in this example could be edited during step 3 in a similar way. Depending on the source, automatic editing could rely more on deductive correction or more on error localization. In general, there may be units for which edit rules (1), (2) and (6) cannot be satisfied simultaneously, e.g., due to errors in $\mathbf{x}_i^{(p)}$ that cannot be solved consistently with the bounds on $\mathbf{y}_i^{(p)}$ derived in step 2. If such a unit is influential, it should be investigated manually instead.

IV. A score to evaluate the suitability of automatic editing across data sources

36. In practice, each data source p may contain some units that exhibit very large inconsistencies between $\mathbf{y}_i^{(p)}$ on one hand and $\left(\mathbf{y}_i^{(1)}, \dots, \mathbf{y}_i^{(p-1)}, \mathbf{y}_i^{(p+1)}, \dots, \mathbf{y}_i^{(P)}\right)$ on the other hand. During step 3 of the above procedure for automatic editing, these units are likely to require many and/or large adjustments to make the values in $\left(\mathbf{x}_i^{(p)}, \mathbf{y}_i^{(p)}\right)$ simultaneously consistent with all internal edit rules (1) and (2) and the additional edit rules (6) that are derived in step 2. If it is not possible to explain these large inconsistencies by systematic errors that can be resolved confidently using deductive correction rules, they have to be resolved using error localization and the quality of the adjusted data may be low, with many (large) adjustments on the wrong variables.

37. It may therefore be desirable to exclude such units during step 3 of automatic editing across data sources. If the inconsistencies are large enough, they may be edited interactively instead. A local score function that identifies units with large inconsistencies on common variable l may be computed after step 2 as follows:

$$S_{il}^{(p)} = \frac{\left|y_{il}^{(p)} - z_{il}\right|}{|z_{il}|+1}, \quad l = 1, \dots, L. \tag{7}$$

Here, $y_{il}^{(p)}$ denotes an original value (prior to step 1) and z_{il} denotes an imputed 'true' value from step 2. These local scores may be combined into a global score for unit i in source p:

$$S_i^{(p)}(\alpha) = \left\{ \frac{\sum_{l=1}^{L} \left(w_l S_{il}^{(p)}\right)^{\alpha}}{\sum_{l=1}^{L} w_l^{\alpha}} \right\}^{1/\alpha}, \tag{8}$$

where $w_1, \dots, w_L$ are weights that reflect the relative importance of the common variables and $\alpha \geq 1$ is a parameter that determines the shape of the global score function; see Vaasen-Otten et al. (2022) for more details. Units with $S_i^{(p)}(\alpha)$ above some threshold may be excluded from automatic editing of source p in step 3.

V. Conclusions

38. At Statistics Netherlands, we have conducted an extensive Proof of Concept (PoC) with automatic editing across data sources, using the method that was outlined in Section III. Different settings were tried and the results were discussed with subject-matter experts associated to the different data sources. A detailed discussion of this PoC is not possible within the limited space of this paper. However, the main conclusions were as follows:

- The proposed stepwise approach to automatic editing is technically feasible within the current IT environment at Statistics Netherlands. It requires much less computational work than an approach that attempts to solve the data editing problem across data sources in one step.
- The use of special edit operations within an extension of the Fellegi-Holt paradigm leads to greater flexibility during error localization. We were able to mimic several commonly-used data editing strategies of subject-matter experts during automatic editing by defining appropriate edit operations.
- To use automatic editing across data sources during regular production, more work is needed to increase the quality of the edited data. This also requires more experience with interactive editing across data sources, to identify and explain possible systematic differences between data sources and to reach consensus among experts about the 'ideal' way to resolve certain types of inconsistencies between data sources. An ongoing process will be needed to learn from experiences during interactive editing within and across data sources and to translate this knowledge into rules, edit operations and other parameter settings for automatic editing.

We are planning to further examine the possibilities for automatic editing across data sources in a pilot for a cluster of statistics, including the annual business statistics.

39. As noted at the beginning of this paper, this research into automatic editing across data sources is part of a larger program at Statistics Netherlands to develop a new integrated uniform production system for business statistics. The main goals of this renewal program are more flexible output, a more agile and efficient production process, and facilitating more possibilities for innovations and further developments.

VI. References

B. Chen, Y. Thibaudeau and W.E. Winkler (2003), A Comparison Study of ACS IF-Then-Else, NIM, DISCRETE Edit and Imputation Systems using ACS Data. UNECE Work Session on Statistical Data Editing, Madrid.

J. Daalmans and S. Scholtus (2018), A MIP Approach for a Generalised Data Editing Problem. Discussion Paper, Statistics Netherlands, The Hague, available at: www.cbs.nl.

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

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

GSDEM (2019), Generic Statistical Data Editing Model, version 2.0. UNECE, available at: statswiki.unece.org/display/sde/GSDEM.

G.E. Liepins (1980), A Rigorous, Systematic Approach to Automatic Data Editing and its Statistical Basis. Report ORNL/TM-7126, Oak Ridge National Laboratory.

J. Pannekoek, S. Scholtus and M. van der Loo (2013), Automated and Manual Data Editing: A View on Process Design and Methodology. *Journal of Official Statistics* **29**, 511–537.

S. Scholtus (2016), A Generalized Fellegi-Holt Paradigm for Automatic Error Localization. *Survey Methodology* **42**, 1–18.

A. Vaasen-Otten, F. Aelen, S. Scholtus and W. de Jong (2022), Towards a New Integrated Uniform Production System for Business Statistics at Statistics Netherlands: Quality Indicators to Guide Top-down Analysis. UNECE Expert Meeting on Statistical Data Editing.

M. van der Loo and E. de Jonge (2018), *Statistical Data Cleaning with Applications in R*. John Wiley & Sons, Hoboken, NJ.