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: quality indicators to guide top-down analysis

Anita Vaasen-Otten, Frank Aelen, Sander Scholtus and Wilco de Jong (Statistics Netherlands, The Netherlands)

amvj.vaasen-otten@cbs.nl; f.aelen@cbs.nl; s.scholtus@cbs.nl; w.dejong@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 processes, 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.

B. Principles of the new production system

3. 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.

(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. This has a number of advantages:

- Unnecessary manual work is avoided;
- Provisional output is available right from the start in order to be able to manage the additional manual work in a top-down manner (see also principle 2);
- Different statistics can be immediately confronted with each other for initial analyses, even if only little input data is available for one of these statistics.

Note that the early compilation of aggregates and draft output for analysis requires microdata to be available right from the start. In the first instance, these can be imputations, for example based on t-1, register data, or other reference data. Once a response record is available, it replaces the previous imputations;

(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. Then selective interactive editing is applied. The selection of units for interactive editing is controlled by the effect on the quality of the output. This means that only units with suspicious values and high impact are selected. Additional manual work takes place within the Large Case Unit for the large and very complex businesses;

(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 from various statistics and other sources are therefore confronted as early as possible in the process. This confrontation of data takes place within clusters of related statistics. Clusters are formed on the basis of common characteristics such as population or theme. First, for each individual business, all available data within the cluster is automatically edited in conjunction where possible. Next, using selective interactive editing, the output requirements of the individual statistics are leading in further making the data consistent. An additional advantage of making data consistent at the micro level early on is that flexible output can be created with better quality. Further integration of the data across the clusters of statistics takes place later on in the process (at National Accounts);

(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. For example, everyone within a cluster of statistics only has access to all data of all statistics within that cluster. Even if the data trickles in, it is made available immediately. This also applies to edited microdata and the aggregates and provisional output composed from this. By making this data immediately available, different statistics can be confronted with each other at an early stage (see also principle 3);

(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. It can occur that data cannot be linked;

(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. In this way optimal reuse of data is facilitated;
- We work with coordinated statistical units. This makes it possible to respond more flexibly to phenomenon-oriented output involving many data sources. Certain units are necessary for individual statistical requirements; these are also coordinated centrally;
- 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. We centrally manage and support processes, methods, data and IT, so that optimal reuse can be made of these building blocks;

(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

4. The main focus of this paper is the quality indicators we use in the new production system to guide the top-down analysis. For this, in section II score functions are described for use in single statistics as well as for confrontations between statistics. Section III gives some examples of dashboards that were developed in a Proof of Concept, incorporating the score functions. Finally, in section IV the main conclusions follow. Note that automatic data editing is described in a companion paper by the same authors.

II. Overview of score functions

A. Score functions

5. In this section quality indicators are described that guide the top-down data analysis in order to optimize the limited amount of manual data editing. For an overview of this topic in general, see, e.g., Granquist (1995), Granquist and Kovar (1997) and De Waal et al. (2011). We focus on so-called score functions that identify potential influential errors in the data; see also, e.g., Lawrence and McKenzie (2000). Below, the essence of score functions is briefly explained, limited to the score functions that we have already implemented in Proofs of Concept or in production. More details and other score functions are available in internal Statistics Netherlands papers, written in Dutch.

6. Scores can relate to different levels. The score for an entire record (unit) is also called a global score, which in turn is made up of several local scores (Hedlin, 2008; De Waal et al., 2011). These local scores can concern, for example, the plausibility of the value of an individual variable within the record, the ratio or structure of multiple variables within the record or even a comparison of a certain core variable with another source or statistic. The prioritization of records for manual editing is based on the global scores. The local scores can then be used to quickly zoom in on the part of the record where there may be a problem. In addition, a measure can be derived from all the global scores of the records in an output aggregate, to indicate the expected quality of that aggregate.

B. Local scores

7. To make things a little more concrete, here is an example of a local score function s, in this case for a single level variable:

$$s_{i,j} = \frac{v_i * |y_{i,j} - \tilde{y}_{i,j}|}{|Y_j|}, \quad (j = 1, \dots, J) \tag{1}$$

with v_i the sample weight of unit i (e.g. an enterprise), $y_{i,j}$ the observed value of variable j, $\tilde{y}_{i,j}$ a reference value for variable j and Y_j an estimate for the aggregate total for variable j. This simple example demonstrates how the principle of a score function works. The score determines the difference between an observed value of, for example, turnover or personnel costs, with an expected reference value, taking the sample weight of the enterprise into account. This difference is divided by the total value of the variable at publication level, for example the Manufacturing Industry or a part thereof. This gives the relative influence of the possible error on the output. There are several options for determining the reference values, such as values from t-1, from another source or a related variable, possibly using a regression or ratio estimator or corrected with a development. Also, a robust representative value (for example, a median) can be taken of similar units. Note that the index $j = 1, \dots, J$ runs over the local scores, not the target variables. In general, the same target variable can appear in multiple local scores.

8. Depending on the specific application, other local score functions can also be defined. For example, for structure variables in the Structural Business Statistics (SBS), we use local scores of the form:

$$s_{i,j} = \frac{|v_i*(y_{i,j}-\tilde{q}_{i,j}*o_{i,j})|}{|z_j|}, \quad (j = 1, \dots, J) \tag{2}$$

where $o_{i,j}$ is the value of unit i for a variable o_j (usually total revenues for the SBS) to which the target variable $y_{i,j}$ (e.g. a sub-variable that is strongly correlated with total revenues) is compared. Furthermore, $\tilde{q}_{i,j}$ is a reference value for the ratio $q_{i,j} = y_{i,j}/o_{i,j}$, based e.g. on a previous period or similar units. The denominator contains either $Z_j = Y_j$ (for target variables with strictly positive totals), or $Z_j = O_j$ (for target variables for which a total can be ≤ 0). In this case, analysis and data editing of the variable o_j, should be done before analysis of the sub-variables $y_{i,j}$, so that the variable o_j can be regarded as fixed.

9. Besides (1) and (2), more advanced local score functions can be derived, in which two or more target variables occur simultaneously. We used these e.g. in the SBS to deal with relationships between variables that are important for national accounts, such as developments of production-use ratios in constant prices, or labor costs per full-time equivalent (FTE). Dealing with multiple target variables required the application of Taylor linearization. More details are available in an internal Statistics Netherlands paper, written in Dutch.

C. Global scores

10. Ultimately, we want one global score per unit that is compiled from the underlying local scores, because when editing manually, an analyst usually treats the entire record of a unit and not just one variable. It helps that the local scores are always expressed as potential relative effects on the outcomes (dividing, for example, by the aggregate total). This means that the local scores are dimensionless and can be added up in a weighted manner to give a global score s_i for unit i:

$$s_i(\alpha) = \left\{ \frac{\sum_{j=1}^{J} \left(w_j * \frac{s_{i,j}}{M_j} \right)^{\alpha}}{\sum_{j=1}^{J} w_j^{\alpha}} \right\}^{1/\alpha} . \tag{3}$$

Adjustable weights w_j can be used to indicate that certain target variables, such as totals of revenue or costs, are more important than others (such as their details). M_j is a measure for the 'maximum acceptable' relative influence per unit of a possible error in the target variable(s) of local scores $s_{i,j}$ on the aggregate in the denominator and partly determines the influence of the local score on a global score. Larger values of M_j could be assigned to local scores, for instance, when larger deviations from reference values are expected naturally because the underlying variables are more volatile. Taking such differences into account, the magnitude of the normalized local score $s_{i,j}/M_j$ should have the same interpretation for all j. Values for w_j and M_j may be set differently for different output aggregates; default values are 1. The parameter α determines the shape of the global score function. It is required that $\alpha \geq 1$; usually $\alpha = 1$ (weighted average), $\alpha = 2$ (weighted Euclidean norm) or $\alpha = \infty$ (weighted maximum) is chosen (Hedlin, 2008).

11. Similar to global scores, other composite scores can be determined for a block of related variables on a questionnaire, or for multiple local scores belonging to the same variable.

D. Aggregate scores

12. Furthermore, a summary measure is calculated based on the scores for all units that contribute to a particular output aggregate A. Since it is particularly interesting whether there are still (many) scores with high values within an aggregate, the aggregate score only counts scores above a certain threshold τ_A. This can be done for local scores:

$$S_{j,A} = \sum_{i \in A} \frac{s_{i,j}}{M_j} * I\left\{ \frac{s_{i,j}}{M_j} \geq \tau_A \right\}, \quad (j = 1, \dots, J), \tag{4}$$

and for global (or other composite) scores:

$$S_A(\alpha) = \sum_{i\in A} s_i(\alpha) * I\{s_i(\alpha) \geq \tau_A\}, \tag{5}$$

where $I\{.\} = 1$ if the argument is true and else $I\{.\} = 0$. The specific values of the threshold can be based on past experience or can be determined based on e.g. sample variances or the required accuracy of the output aggregates. Regardless of the exact threshold value, this provides us with a simple measure that can help to prioritize which output aggregates still need the most attention, thus contributing to an efficient production process.

E. Score functions for confrontations between statistics

13. Top-down analysis will be performed not only for individual statistics, but also in an integrated manner for clusters of related statistics with one or more overlapping variables. This will result in more consistency across statistics and with national accounts and will avoid situations in which a specific event for a particular business is investigated for each statistic separately. With these aims in mind, we have developed particular score functions for top-down analysis that identify potential influential errors with respect to consistency between statistics and with respect to economic indicators for national accounts. As with the individual statistics, we also use local, global and aggregate scores for confrontations between statistics.

14. For each of the (overlapping) target variables y_l $(l = 1, \ldots, L)$ for which a confrontation between the statistics takes place (confrontation variable), we denote the observed values for unit i in the different statistics (or other sources) as $y_{il}^{(1)}, \ldots, y_{il}^{(P)}$, with P the number of confronted statistics. The local score function for the confrontation of statistic p with statistic q for this variable has the following form:

$$s_{il}^{(p,q)} = \frac{\left|v_i^{(p)} * \left(y_{il}^{(p)} - y_{il}^{(q)}\right)\right|}{|Y_l|}, \quad (l = 1, \ldots, L; p = 1, \ldots, P; q = 1, \ldots, P; q \neq p) \tag{6}$$

Here $v_i^{(p)}$ is the sample weight of unit i for statistic p. The aggregate Y_l is a robust estimate based on the different aggregates for the respective target variable from the separate statistics, $Y_l^{(p)} = \sum_i v_i^{(p)} * y_{il}^{(p)}$ for $p = 1, \ldots, P$. If any of the statistics has been edited before, it could be used as a robust estimate. In other situations, for example, the median of the aggregates $Y_{l,G}^{(1)}, \ldots, Y_{l,G}^{(P)}$ could be chosen, or an adjusted estimate in which possible outliers have a reduced weight. For confrontations between statistics, the aggregates of the national accounts supply-use tables will often be chosen as the aggregation level. Note that in general $s_{il}^{(p,q)} \neq s_{il}^{(q,p)}$. Furthermore, P can depend on l and i, but this is not explicitly stated here in order not to complicate the notation unnecessarily.

15. Aggregate scores for the confrontation between statistics can be derived in the same manner as in (4):

$$S_{l,A}^{(p,q)} = \sum_{i\in A} \frac{s_{il}^{(p,q)}}{M_l^{(p,q)}} * I\left\{\frac{s_{il}^{(p,q)}}{M_l^{(p,q)}} \geq \tau_A\right\}, \quad (l = 1, \ldots, L; p = 1, \ldots, P_0; q = 1, \ldots, P; q \neq p). \tag{7}$$

16. Similarly, several other scores can be derived, such as global scores per unit and confrontation variables across all statistics, global scores per unit across all confrontation variables and statistics or global scores per pair of statistics across all relevant variables, plus the associated aggregate scores. These we have not yet implemented.

F. Score functions regarding population dynamics

17. The business population is constantly changing: businesses are founded, change in size and composition and discontinue. This has consequences for economic statistics. In many statistical processes time is spent on population dynamics during analysis. How population dynamics are handled often differs per statistic. In a Proof a Concept (POC) we investigated whether it is possible to simplify the analysis of population dynamics and to

avoid duplication of work in statistics. For this, a scoring method has been developed, similar to the scores described above, and in addition detailed information is made available in a dashboard about unit events. When implemented in production, information about how population dynamics are handled in specific cases can also be exchanged between statistics.

18. In top-down analysis there is need for a score that measures the influence of certain population dynamics and that can prioritize whether further attention is needed from an analyst. In the POC we have limited ourselves to business formations and liquidations. Extensions to other unit events will have to be investigated later. As a variable used to calculate the influence of such an event, we started with the number of employees per unit ($x_{i,t}$), because this information is available for all business units from tax data. Scores are calculated based on population dynamics from period t-1 to period t. A relatively simple example of a score function for population dynamics can be found using equation (1) with $v_i = 1$, $y_i = \delta_{i,t,A} * x_{i,t}$, $Y = X_{A,t}$ and $\tilde{y}_i = \delta_{i,t-1,A} * x_{i,t-1} * X_{A,t}/X_{A,t-1}$, where $\delta_{i,t,A}$ indicates whether unit i belongs to aggregate A in time period t. More details of these and other scores for population dynamics are available in an internal Statistics Netherlands paper, written in Dutch.

III. Use of score functions in top-down analysis

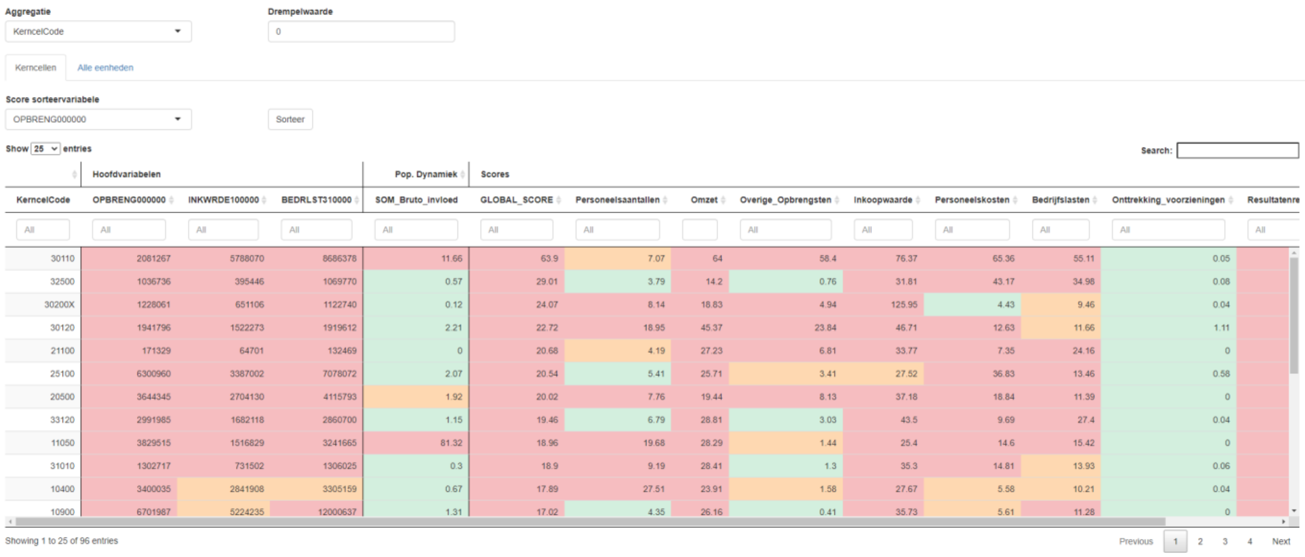
A. Display of score functions on dashboards for top-down analysis

19. The different scores described above, are displayed on dashboards for top-down analysis. Using the scores, and possibly other indicators like e.g. response rates, it is possible to zoom in from e.g. NACE sections to NACE subdivisions to unit records to suspicious variables in a record. Response and edited values of the variables of the different statistics are color-coded according to their respective scores (at the moment green, orange and red for low, medium and high scores, respectively). This enables an efficient analysis of large amounts of data.

20. Below, some examples are given of dashboards that were developed in a Proof of Concept (POC). In this POC the lay-out of the dashboards is focused on the use of score functions. Other aspects may be added for specific implementations in production. Note that data from enterprises that belong to the Large Case Unit (top 360 enterprise groups) are excluded in this analysis, because they are dealt with by specialized analysts.

B. Examples of the use of score functions within a single statistic

21. Dashboard with overview of publication aggregates ("KerncelCode") of the annual Production Statistics (part of SBS), with some key variables, scores for population dynamics, global scores and composite scores for the different variable blocks. The displayed values are aggregate scores as defined in (4) and (5). This concerns data that has only been edited automatically and which no analyst has yet edited manually, hence the many red cells (i.e. high scores), because there are potential influential errors in almost every aggregate in that situation.



Aggregatie: KerncelCode
Drempelwaarde: 0

Kerncellen | Alle eenheden

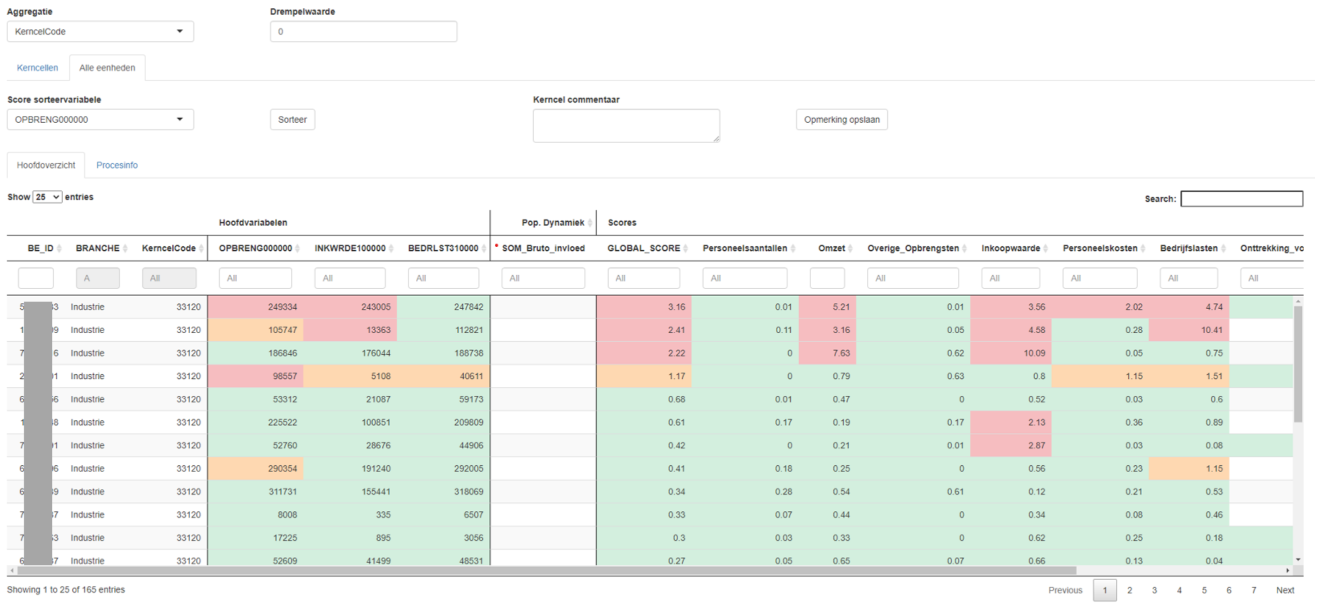
Score sorteervariabele: OPBRENG000000
Sorteer

Show 25 entries
Search:

	Hoofdvariabelen			Pop. Dynamiek	Scores								
KerncelCode	OPBRENG000000	INKWRDE100000	BEDRLST310000	SOM_Bruto_invloed	GLOBAL_SCORE	Personeelsaantallen	Omzet	Overige_Opbrengsten	Inkoopwaarde	Personeelskosten	Bedrijfslasten	Onttrekking_voorzieningen	Resultatenre
All	All	All	All	All	All	All		All	All	All	All	All	All
30110	2081267	5788070	8686378	11.66	63.9	7.07	64	58.4	76.37	65.36	55.11	0.05	
32500	1036736	395446	1069770	0.57	29.01	3.79	14.2	0.76	31.81	43.17	34.98	0.08	
30200X	1228061	651106	1122740	0.12	24.07	8.14	18.83	4.94	125.95	4.43	9.46	0.04	
30120	1941796	1522273	1919612	2.21	22.72	18.95	45.37	23.84	46.71	12.63	11.66	1.11	
21100	171329	64701	132469	0	20.68	4.19	27.23	6.81	33.77	7.35	24.16	0	
25100	6300960	3387002	7078072	2.07	20.54	5.41	25.71	3.41	27.52	36.83	13.46	0.58	
20500	3644345	2704130	4115793	1.92	20.02	7.76	19.44	8.13	37.18	18.84	11.39	0	
33120	2991985	1682118	2860700	1.15	19.46	6.79	28.81	3.03	43.5	9.69	27.4	0.04	
11050	3829515	1516829	3241665	81.32	18.96	19.68	28.29	1.44	25.4	14.6	15.42	0	
31010	1302717	731502	1306025	0.3	18.9	9.19	28.41	1.3	35.3	14.81	13.93	0.06	
10400	3400035	2841908	3305159	0.67	17.89	27.51	23.91	1.58	27.67	5.58	10.21	0.04	
10900	6701987	5224235	12000637	1.31	17.02	4.35	26.16	0.41	35.73	5.61	11.28	0	

Showing 1 to 25 of 96 entries

22. Dashboard with an overview of the units (“BE_ID”) within a publication aggregate to be analyzed. The colors orange and red indicate which units should be looked at by an analyst. It can also already be seen in which key variable and/or variable block the influential suspect values are located.



Aggregatie: KerncelCode

Drempelwaarde: 0

Score sorteervariabele: OPBRENG000000

Kerncel commentaar:

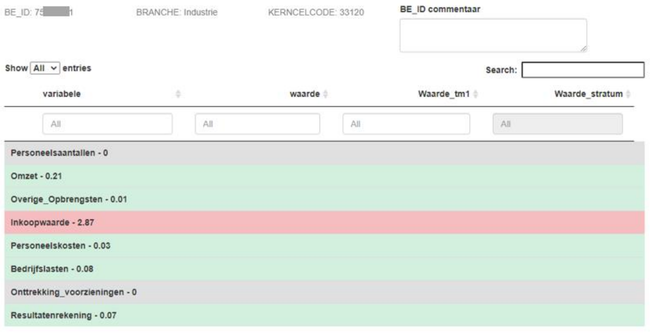
Show 25 entries

Search:

			Hoofdvariabelen			Pop. Dynamiek	Scores							
BE_ID	BRANCHE	KerncelCode	OPBRENG000000	INKWRDE100000	BEDRLST310000	• SOM_Bruto_invloed	GLOBAL_SCORE	Personeelsaantallen	Omzet	Overige_Opbrengsten	Inkoopwaarde	Personeelskosten	Bedrijfslasten	Onttrekking_vo
	A	All	All	All	All	All	All	All		All	All	All	All	All
5[redacted]3	Industrie	33120	249334	243005	247842		3.16	0.01	5.21	0.01	3.56	2.02	4.74	
1[redacted]9	Industrie	33120	105747	13363	112821		2.41	0.11	3.16	0.05	4.58	0.28	10.41	
7[redacted]6	Industrie	33120	186846	176044	188738		2.22	0	7.63	0.62	10.09	0.05	0.75	
2[redacted]1	Industrie	33120	98557	5108	40611		1.17	0	0.79	0.63	0.8	1.15	1.51	
6[redacted]6	Industrie	33120	53312	21087	59173		0.68	0.01	0.47	0	0.52	0.03	0.6	
1[redacted]8	Industrie	33120	225522	100851	209809		0.61	0.17	0.19	0.17	2.13	0.36	0.89	
7[redacted]1	Industrie	33120	52760	28676	44906		0.42	0	0.21	0.01	2.87	0.03	0.08	
6[redacted]6	Industrie	33120	290354	191240	292005		0.41	0.18	0.25	0	0.56	0.23	1.15	
6[redacted]9	Industrie	33120	311731	155441	318069		0.34	0.28	0.54	0.61	0.12	0.21	0.53	
7[redacted]7	Industrie	33120	8008	335	6507		0.33	0.07	0.44	0	0.34	0.08	0.46	
7[redacted]3	Industrie	33120	17225	895	3056		0.3	0.03	0.33	0	0.62	0.25	0.18	
6[redacted]7	Industrie	33120	52609	41499	48531		0.27	0.05	0.65	0.07	0.66	0.13	0.04	

Showing 1 to 25 of 165 entries

23. Left: the editing screen for an individual unit. In this case, it can be clearly seen that the influential suspect values are in the Purchase Value (“Inkoopwaarde”) block. Right: same screen, but with the Purchase Value block unfolded. Values for t (“waarde”) and t-1 (“Waarde_tm1”) or stratum means (“Waarde_stratum”) are given for the underlying variables.



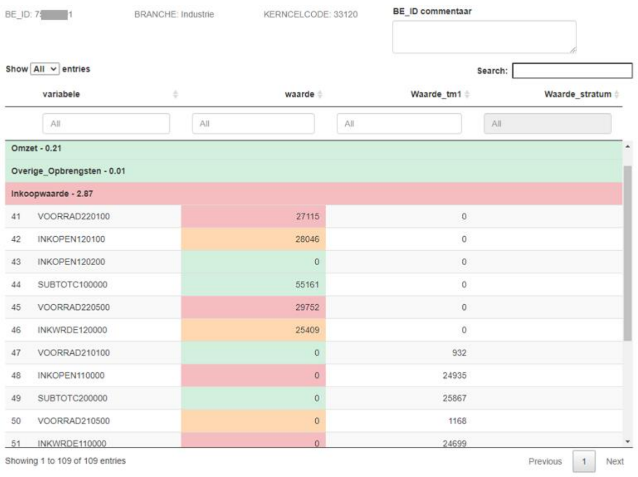
BE_ID: 75[redacted]1 BRANCHE: Industrie KERNCELCODE: 33120

BE_ID commentaar:

Show All entries

Search:

variabele	waarde	Waarde_tm1	Waarde_stratum
All	All	All	All
Personeelsaantallen - 0			
Omzet - 0.21			
Overige_Opbrengsten - 0.01			
Inkoopwaarde - 2.87			
Personeelskosten - 0.03			
Bedrijfslasten - 0.08			
Onttrekking_voorzieningen - 0			
Resultatenrekening - 0.07			



BE_ID: 7[redacted]1 BRANCHE: Industrie KERNCELCODE: 33120

BE_ID commentaar:

Show All entries

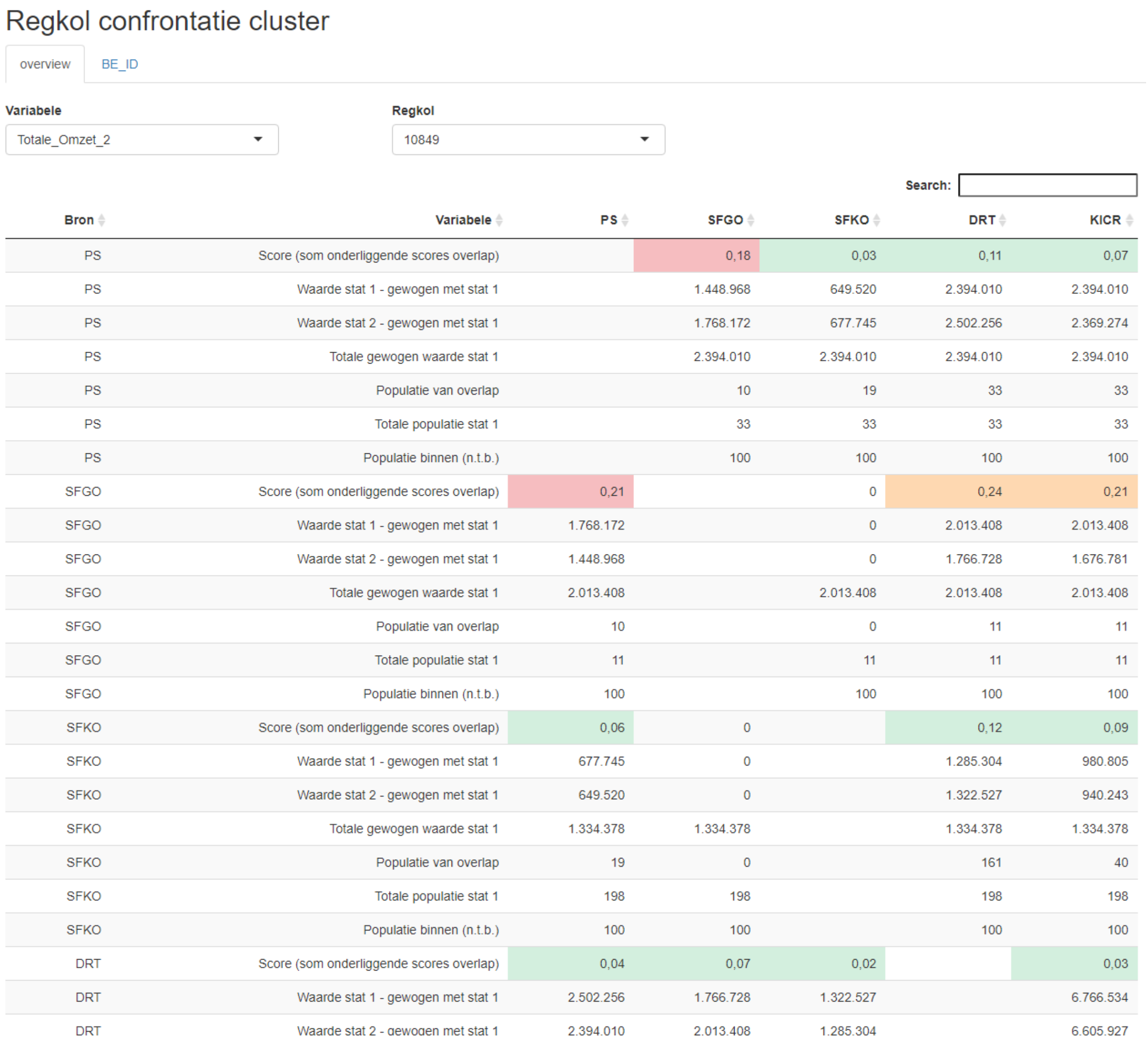
Search:

	variabele	waarde	Waarde_tm1	Waarde_stratum
	All	All	All	All
Omzet - 0.21				
Overige_Opbrengsten - 0.01				
Inkoopwaarde - 2.87				
41	VOORRAD220100	27115	0	
42	INKOPEN120100	28046	0	
43	INKOPEN120200	0	0	
44	SUBTOTC100000	55161	0	
45	VOORRAD220500	29752	0	
46	INKWRDE120000	25409	0	
47	VOORRAD210100	0	932	
48	INKOPEN110000	0	24935	
49	SUBTOTC200000	0	25867	
50	VOORRAD210500	0	1168	
51	INKWRDE110000	0	24699	

Showing 1 to 109 of 109 entries

C. Examples of the use of score functions across statistics

24. Dashboard with an overview of influential inconsistencies between different statistics, in this case for the turnover variable for an aggregate that is also used in the national accounts supply-use tables. Statistics that are compared for the turnover variable are the annual Production Statistics (“PS”, part of SBS), Statistics of Finances of Enterprises (for large units based on a survey: “SFGO”; for smaller units based on tax information: “SFKO”) and Short Term Statistics (survey: “KICR”, tax information: “DRT”). Other variables are also compared with the PRODCOM and the Employment and Wages Statistics (“SWL”).



Regkol confrontatie cluster

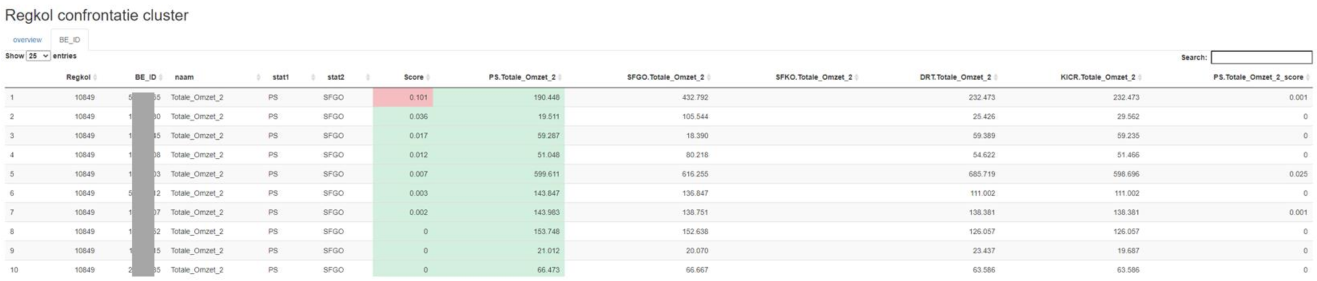
Variabele: Totale_Omzet_2

Regkol: 10849

Search:

Bron	Variabele	PS	SFGO	SFKO	DRT	KICR
PS	Score (som onderliggende scores overlap)		0,18	0,03	0,11	0,07
PS	Waarde stat 1 - gewogen met stat 1		1.448.968	649.520	2.394.010	2.394.010
PS	Waarde stat 2 - gewogen met stat 1		1.768.172	677.745	2.502.256	2.369.274
PS	Totale gewogen waarde stat 1		2.394.010	2.394.010	2.394.010	2.394.010
PS	Populatie van overlap		10	19	33	33
PS	Totale populatie stat 1		33	33	33	33
PS	Populatie binnen (n.t.b.)		100	100	100	100
SFGO	Score (som onderliggende scores overlap)	0,21		0	0,24	0,21
SFGO	Waarde stat 1 - gewogen met stat 1	1.768.172		0	2.013.408	2.013.408
SFGO	Waarde stat 2 - gewogen met stat 1	1.448.968		0	1.766.728	1.676.781
SFGO	Totale gewogen waarde stat 1	2.013.408		2.013.408	2.013.408	2.013.408
SFGO	Populatie van overlap	10		0	11	11
SFGO	Totale populatie stat 1	11		11	11	11
SFGO	Populatie binnen (n.t.b.)	100		100	100	100
SFKO	Score (som onderliggende scores overlap)	0,06	0		0,12	0,09
SFKO	Waarde stat 1 - gewogen met stat 1	677.745	0		1.285.304	980.805
SFKO	Waarde stat 2 - gewogen met stat 1	649.520	0		1.322.527	940.243
SFKO	Totale gewogen waarde stat 1	1.334.378	1.334.378		1.334.378	1.334.378
SFKO	Populatie van overlap	19	0		161	40
SFKO	Totale populatie stat 1	198	198		198	198
SFKO	Populatie binnen (n.t.b.)	100	100		100	100
DRT	Score (som onderliggende scores overlap)	0,04	0,07	0,02		0,03
DRT	Waarde stat 1 - gewogen met stat 1	2.502.256	1.766.728	1.322.527		6.766.534
DRT	Waarde stat 2 - gewogen met stat 1	2.394.010	2.013.408	1.285.304		6.605.927

25. Overview of the underlying units of the two suspicious red cells from the previous dashboard, both related to a large inconsistency between PS and SFGO for the turnover variable. This shows which units are causing the inconsistency. The green colors at "PS.Totale_Omzet_2" indicate that there is no suspicious development in any unit relative to t-1, if you look purely at the PS. Clicking on the red cell for the unit with a large individual inconsistency between PS and SFGO, gives more detailed information in the underlying editing dashboards as shown before.



Regkol confrontatie cluster

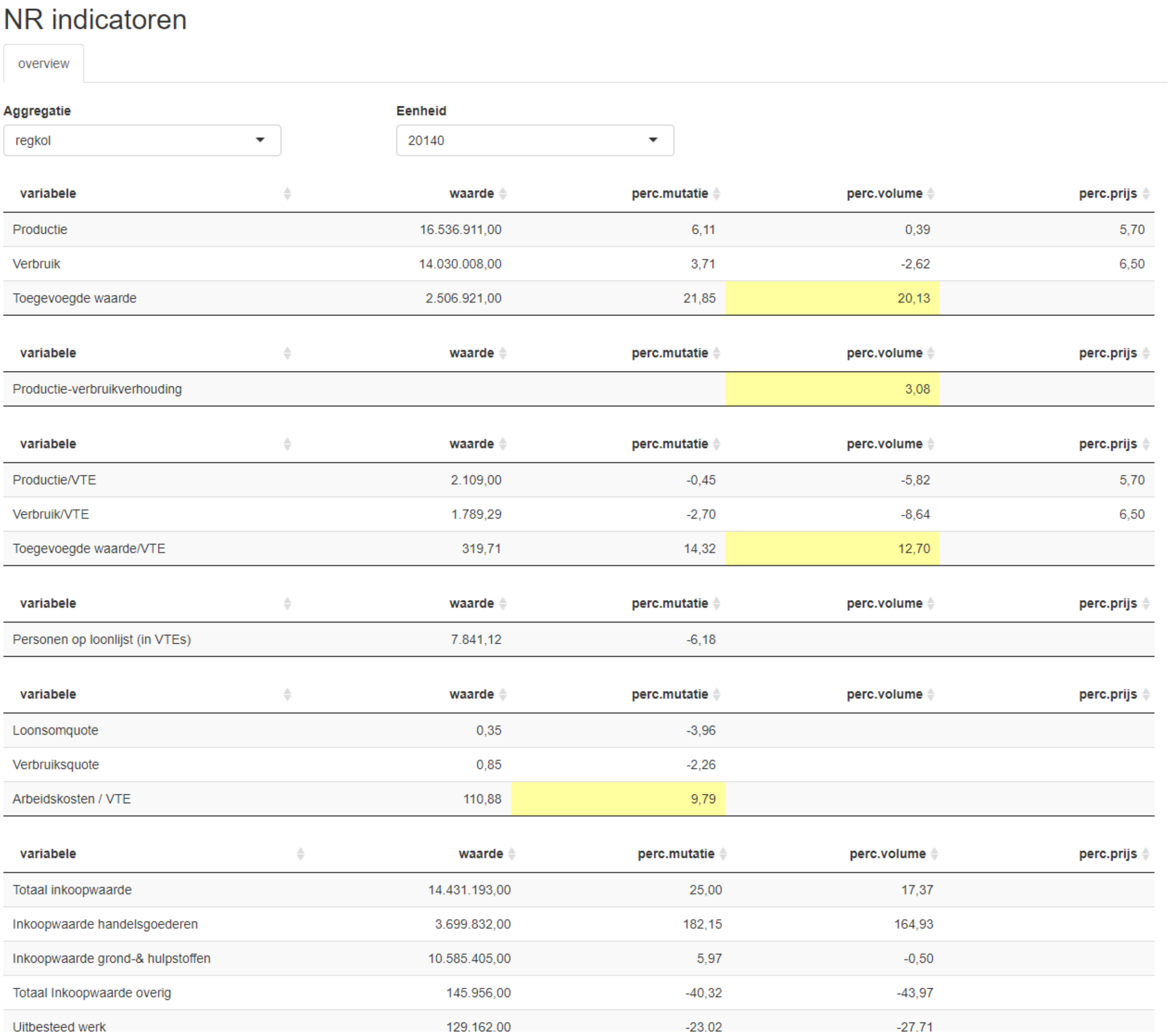
Show 25 entries

Search:

	Regkol	BE_ID	naam	stat1	stat2	Score	PS.Totale_Omzet_2	SFGO.Totale_Omzet_2	SFKO.Totale_Omzet_2	DRT.Totale_Omzet_2	KICR.Totale_Omzet_2	PS.Totale_Omzet_2_score
1	10849	[redacted]	Totale_Omzet_2	PS	SFGO	0.101	190.448	432.792		232.473	232.473	0.001
2	10849	[redacted]	Totale_Omzet_2	PS	SFGO	0.036	19.511	105.544		25.426	29.562	0
3	10849	[redacted]	Totale_Omzet_2	PS	SFGO	0.017	59.287	18.390		59.389	59.235	0
4	10849	[redacted]	Totale_Omzet_2	PS	SFGO	0.012	51.048	80.218		54.622	51.466	0
5	10849	[redacted]	Totale_Omzet_2	PS	SFGO	0.007	599.611	616.255		685.719	598.696	0.025
6	10849	[redacted]	Totale_Omzet_2	PS	SFGO	0.003	143.847	136.847		111.002	111.002	0
7	10849	[redacted]	Totale_Omzet_2	PS	SFGO	0.002	143.983	138.751		138.381	138.381	0.001
8	10849	[redacted]	Totale_Omzet_2	PS	SFGO	0	153.748	152.638		126.057	126.057	0
9	10849	[redacted]	Totale_Omzet_2	PS	SFGO	0	21.012	20.070		23.437	19.687	0
10	10849	[redacted]	Totale_Omzet_2	PS	SFGO	0	66.473	66.667		63.586	63.586	0

D. Examples of the use of more complex score functions, e.g. for national accounts

26. Dashboard with an overview of indicators for national accounts, based on SBS variables. Aggregates are used that are relevant for national accounts. The four yellow indicated variables have the focus, from top to bottom: developments of value added, production-use ratio, value added per FTE, all in constant prices, and labor costs per FTE. Analysts can estimate whether the values deviate from what was expected and then investigate the underlying units that contribute most to suspicious values according to a score function. In this way, analysts gain insight into the consequences for national accounts before completing the analysis of the SBS and have the opportunity to correct values of influential suspect units.



NR indicatoren

Aggregatie: regkol

Eenheid: 20140

variabele	waarde	perc.mutatie	perc.volume	perc.prijs
Productie	16.536.911,00	6,11	0,39	5,70
Verbruik	14.030.008,00	3,71	-2,62	6,50
Toegevoegde waarde	2.506.921,00	21,85	20,13	

variabele	waarde	perc.mutatie	perc.volume	perc.prijs
Productie-verbruikverhouding			3,08	

variabele	waarde	perc.mutatie	perc.volume	perc.prijs
Productie/VTE	2.109,00	-0,45	-5,82	5,70
Verbruik/VTE	1.789,29	-2,70	-8,64	6,50
Toegevoegde waarde/VTE	319,71	14,32	12,70	

variabele	waarde	perc.mutatie	perc.volume	perc.prijs
Personen op loonlijst (in VTEs)	7.841,12	-6,18		

variabele	waarde	perc.mutatie	perc.volume	perc.prijs
Loonsomquote	0,35	-3,96		
Verbruiksquote	0,85	-2,26		
Arbeidskosten / VTE	110,88	9,79		

variabele	waarde	perc.mutatie	perc.volume	perc.prijs
Totaal inkoopwaarde	14.431.193,00	25,00	17,37	
Inkoopwaarde handelsgoederen	3.699.832,00	182,15	164,93	
Inkoopwaarde grond-& hulpstoffen	10.585.405,00	5,97	-0,50	
Totaal Inkoopwaarde overig	145.956,00	-40,32	-43,97	
Uitbesteed werk	129.162,00	-23,02	-27,71	

E. Implementations and future developments

27. Over the past two years, we have tested and refined the above ideas in POC's. Most of the developed scores are implemented in generalized R-modules and can technically be offered to the different agile teams within Statistics Netherlands as a web service to be used in top-down analysis. The scores can be tailored to the various statistics by means of rules. For a limited number of individual statistics, the scores have been implemented and are already being applied in practice. Sometimes this is done by showing the scores in existing dashboards of the statistics and in case the statistics are redesigned, the new dashboards are used. Experiences to date show that the new scores allow analysts to work in a more targeted way than before.

28. Later this year we will start a pilot regarding the top-down analysis of inconsistencies between statistics. We plan to apply the methods described above, especially the score functions across statistics, on live production data during the regular analysis of the statistics concerned. In this way we can also gain experience with new roles that are necessary for this new way of working. For example, we will experiment with specialized analysts

who focus on the inconsistencies between statistics and learn how these analysts can best work together with the regular analysts of the individual statistics.

IV. Conclusions

29. 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. Important aspects of the new production system are generalized modular building blocks with customizable settings (also reducing IT legacy burden) and the use of quality indicators to guide the top-down data analysis and to optimize the limited amount of manual data editing (in addition to automatic data editing). Top-down analysis will be performed not only for individual statistics, but also in an integrated manner for clusters of related statistics. This will result in more consistency across statistics and with national accounts and will avoid situations in which a specific event for a particular business is investigated for each statistic separately. With these aims in mind, we have developed particular score functions for top-down analysis that identify potentially influential errors with respect to consistency between statistics and with respect to economic indicators for national accounts. In order to further develop our ideas and to test them on realistic data, we have recently conducted a number of Proofs of Concept. In addition, we have already implemented some of the results on quality indicators and top-down analysis in a number of statistics. In the near future we will continue these stepwise developments and implementations and working in an agile manner, we will keep learning from each further step.

V. References

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

L. Granquist (1995), Improving the Traditional Editing Process. In: *Business Survey Methods* (eds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge, and P.S. Kott), John Wiley & Sons, New York, pp. 385–401.

L. Granquist and J. Kovar (1997), Editing of Survey Data: How Much is Enough? In: *Survey Measurement and Process Quality* (eds. L.E. Lyberg, P. Biemer, M. Collins, E.D. de Leeuw, C. Dippo, N. Schwartz, and D. Trewin), John Wiley & Sons, New York, pp. 415–435.

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

D. Lawrence and R. McKenzie (2000), The General Application of Significance Editing. *Journal of Official Statistics* **16**, pp. 243–253.