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

3-10-2022

EBN 2.x: Renewal program for business statistics at SN

Main goals: towards more efficient, more flexible and futureproof business statistics

Efficient, goal-oriented co-creation

Balance on innovation, re-use and implementation

New process, way of working and matching tooling

2.x: considered small steps

Innovation, co-creation and implementations

Innovation

- 8 principles
- Applied methodology
- The proof of the pudding… is in a Proof of Concept!

Co-creation with Business and with Agile teams

- Re-use of best practices
- As much standardized coherent models & common tools as possible!

Implementations

- Small steps instead of a big bang
- Visible results
- Immediate feedback contributes to continuous improvements

Principles of the new production system*

*: Implementation ongoing

Towards real-time processing

1. We process our input automatically and immediately up to provisional output;
2. We measure quality automatically and thus direct the manual work;

Towards more coherence

3. We make our data consistent as early as possible;
4. We share all our data, right from the start;

Towards more standardization and re-use of best practices

5. We centrally manage all our (population) frames, which are the basis of our statistics;
6. We have fully standardized our processes, methods, data and IT;
7. Our processes, methods, data and IT are modular;

Continuous improvement

8. We resolve manual corrections the following iteration in the standard process.

Quality indicators to guide top-down analysis*

- Focus on score functions that identify potential influential errors in the data;
- Used to prioritize records for manual editing and to quickly zoom in on the part of the record where there may be a problem;
- Can also be used to indicate the expected quality of an aggregate.

***Automatic data editing is described in a companion paper**

Local score for level variable

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

with v_i the sample weight of unit i, $y_{i,j}$ 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

- Relative influence of possible error on the output

- Reference value e.g. t-1, other source, related variable, …

Additional local scores

- Structure variables;
- Consistency across statistics;
- Non-linear indicators with two or more target variables simultaneously (e.g. production-use ratios for national accounts).

More details available in paper

Global scores

Global score s_i per unit that is compiled from the underlying local scores:

$$s_i(\alpha) = \left\{ \frac{\sum_{j=1}^{J} \left(w_j * \frac{s_{i,j}}{M_j} \right)^{\alpha}}{\sum_{j=1}^{J} w_j^{\alpha}} \right\}^{1/\alpha} \tag{2}$$

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.

Aggregate scores

Summary measure based on the scores for all units that contribute to a particular output aggregate A. Only scores are counted above a certain threshold τ_A.

This can be done for local scores:

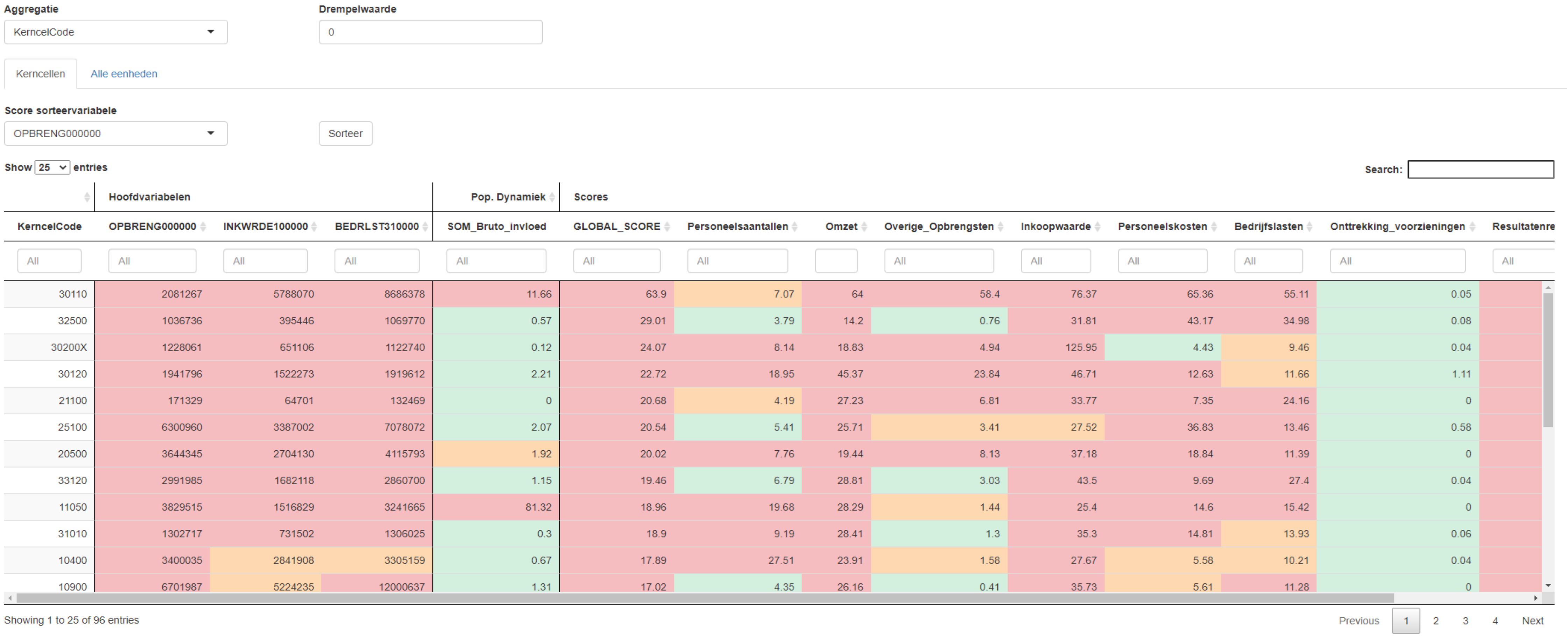
$$S_{j,A} = \sum_{i \in A} \frac{s_{i,j}}{M_j} * I\left\{\frac{s_{i,j}}{M_j} \geq \tau_A\right\}, \quad (j = 1, \ldots, J), \tag{3}$$

and for global (or other composite) scores:

$$S_A(\alpha) = \sum_{i \in A} s_i(\alpha) * I\{s_i(\alpha) \geq \tau_A\}, \tag{4}$$

where $I\{.\} = 1$ if the argument is true and else $I\{.\} = 0$.

Example: single statistic – publication aggregates



Aggregatie: KerncelCode

Drempelwaarde: 0

Score sorteervariabele: OPBRENG000000

Show 25 entries

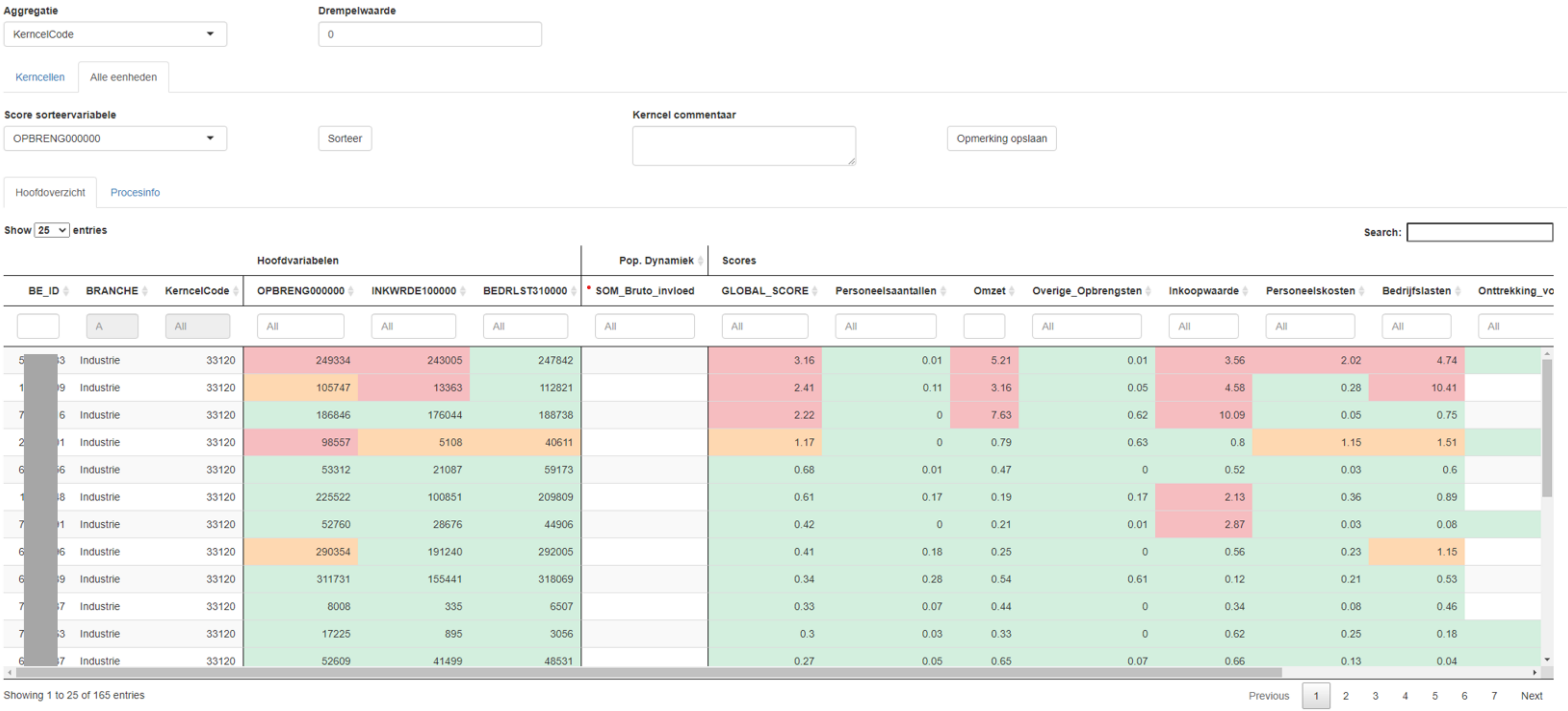
Search:

	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

Prioritize aggregates – based on aggregate score

Example: single statistic – overview of units



Aggregatie
KerncelCode

Drempelwaarde
0

Score sorteervariabele
OPBRENG000000

Kerncel commentaar

Show 25 **entries**

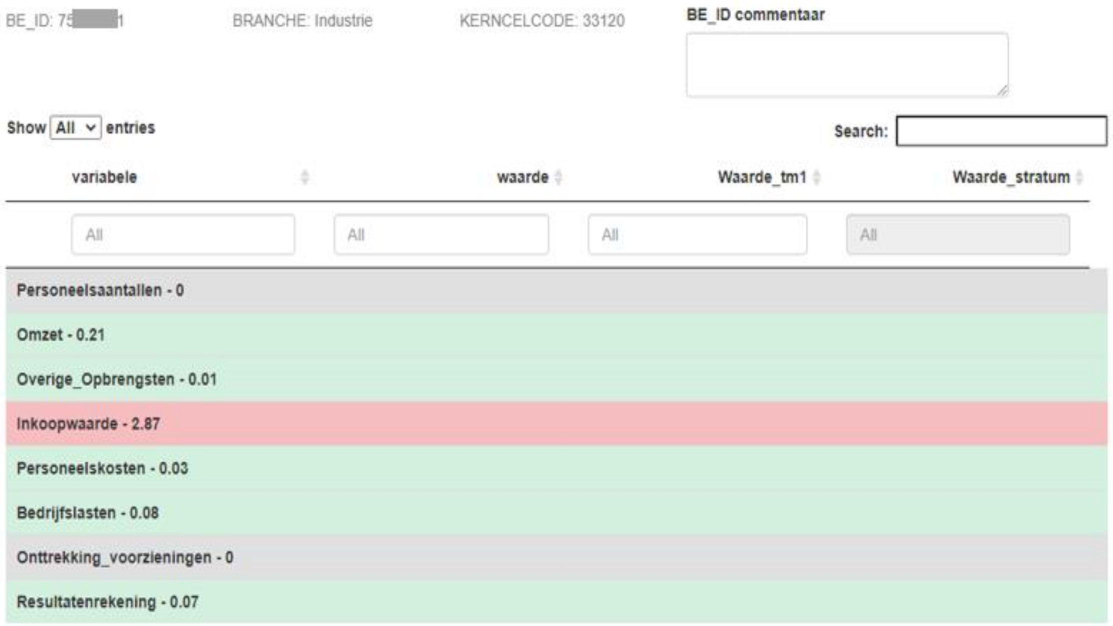
Search:

			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
[redacted]	Industrie	33120	249334	243005	247842		3.16	0.01	5.21	0.01	3.56	2.02	4.74	
[redacted]	Industrie	33120	105747	13363	112821		2.41	0.11	3.16	0.05	4.58	0.28	10.41	
[redacted]	Industrie	33120	186846	176044	188738		2.22	0	7.63	0.62	10.09	0.05	0.75	
[redacted]	Industrie	33120	98557	5108	40611		1.17	0	0.79	0.63	0.8	1.15	1.51	
[redacted]	Industrie	33120	53312	21087	59173		0.68	0.01	0.47	0	0.52	0.03	0.6	
[redacted]	Industrie	33120	225522	100851	209809		0.61	0.17	0.19	0.17	2.13	0.36	0.89	
[redacted]	Industrie	33120	52760	28676	44906		0.42	0	0.21	0.01	2.87	0.03	0.08	
[redacted]	Industrie	33120	290354	191240	292005		0.41	0.18	0.25	0	0.56	0.23	1.15	
[redacted]	Industrie	33120	311731	155441	318069		0.34	0.28	0.54	0.61	0.12	0.21	0.53	
[redacted]	Industrie	33120	8008	335	6507		0.33	0.07	0.44	0	0.34	0.08	0.46	
[redacted]	Industrie	33120	17225	895	3056		0.3	0.03	0.33	0	0.62	0.25	0.18	
[redacted]	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

Prioritize units within aggregate – based on global score

Example: single statistic – individual unit



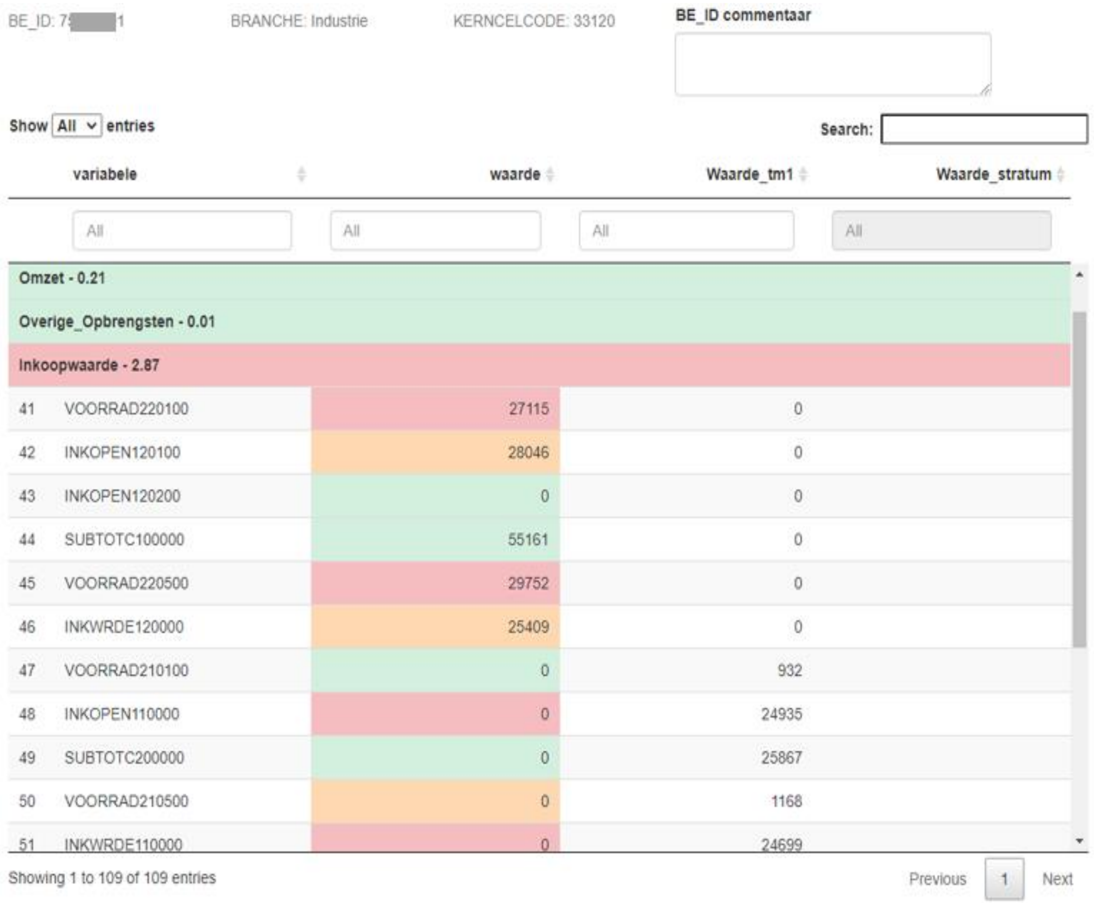
BE_ID: 75[redacted]1 BRANCHE: Industrie KERNCELCODE: 33120

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

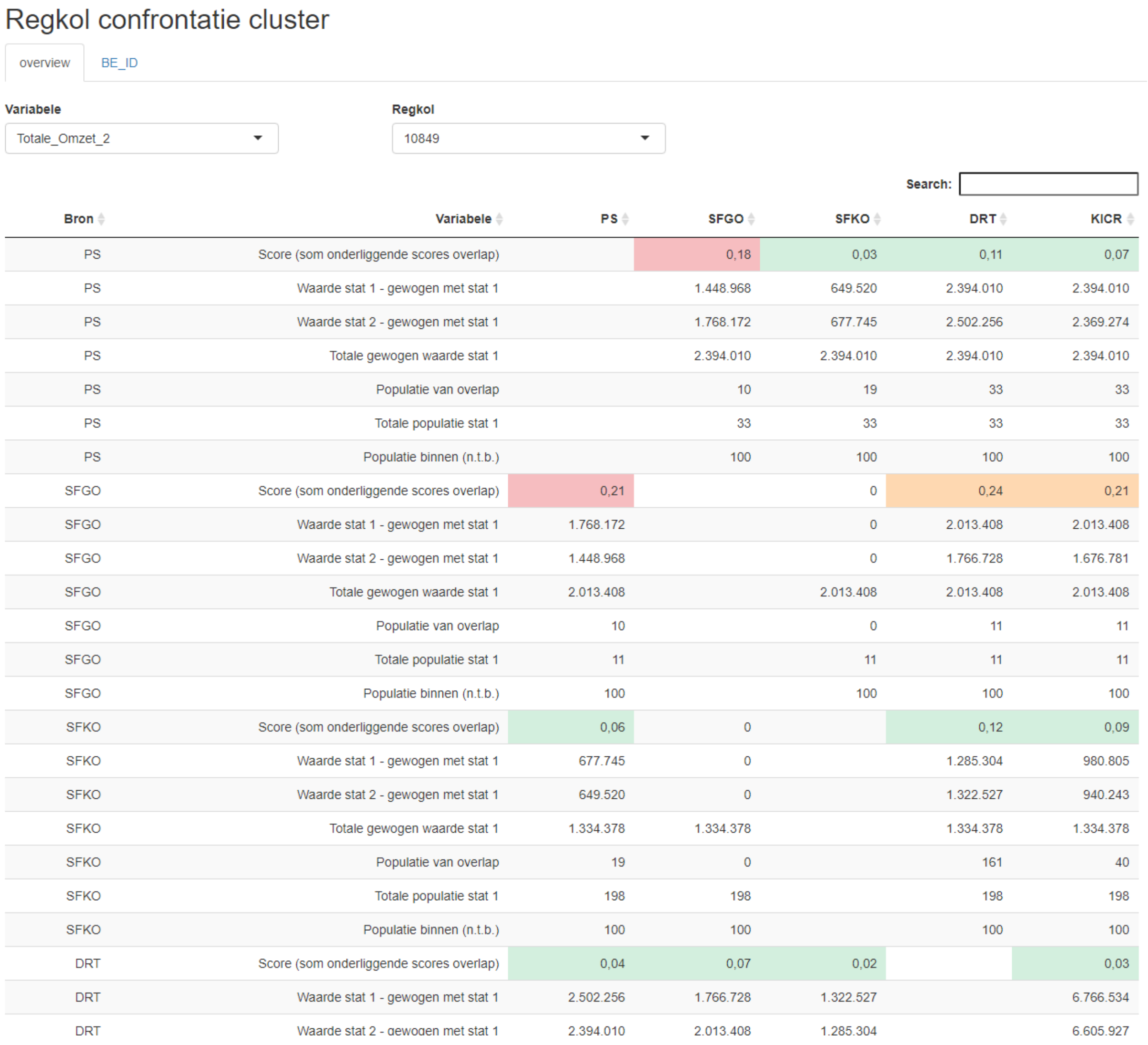
Search:

	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

Prioritize variables within unit – based on local scores

Example: across statistics – publication aggregates



Regkol confrontatie cluster

Variabele: Totale_Omzet_2

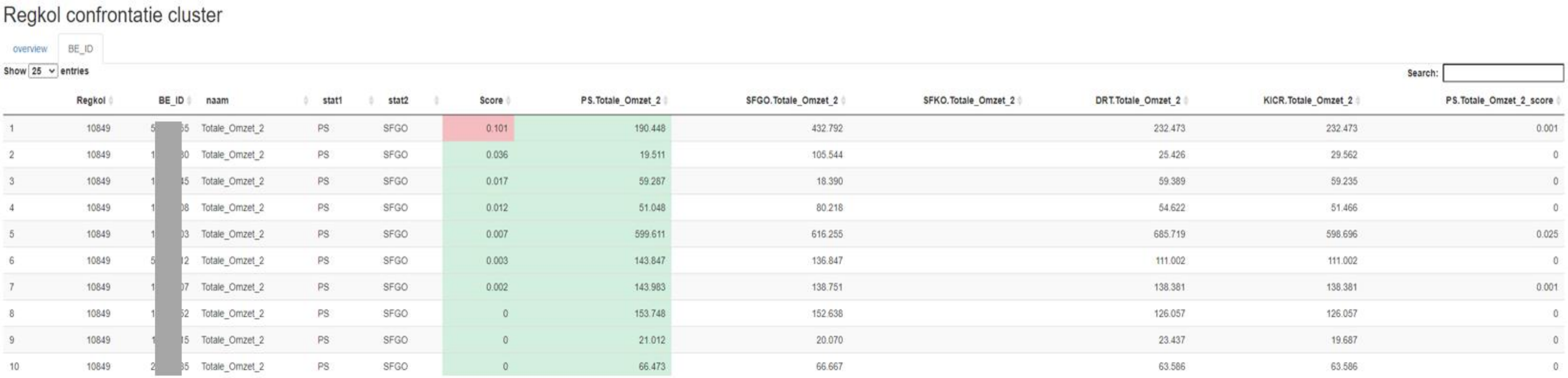
Regkol: 10849

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

Prioritize statistics with large inconsistencies between them, for a certain aggregate and mutual variable

Example: across statistics – overview of units



Regkol confrontatie cluster

	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

Prioritize units within aggregate

Implementations

- Past two years we tested and refined ideas in POCs;
- Implemented in generalized R-modules – web service;
- Scores can be tailored to 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;
- Later this year: pilot regarding the top-down analysis of inconsistencies between statistics → gain experience with new roles that are necessary for this new way of working.

Concluding remarks

- Experiences to date show that the new scores allow analysts to work in a more targeted way than before;
- In the near future we will continue the stepwise developments and implementations and working in an agile manner, we will keep learning from each further step.