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*

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.

*: Implementation ongoing
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

\[ S_{i,j} = \frac{v_i \cdot |y_{i,j} - \tilde{y}_{i,j}|}{|Y_j|}, \quad (j = 1, \ldots, J) \]  

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:

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

(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 $\tau_A$.

This can be done for local scores:

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

(3)

and for global (or other composite) scores:

$$S_A(\alpha) = \sum_{i \in A} s_i(\alpha) \cdot I \left\{ s_i(\alpha) \geq \tau_A \right\},$$

(4)

where $I\{.\} = 1$ if the argument is true and else $I\{.\} = 0$. 
Example: single statistic – publication aggregates

Prioritize aggregates – based on aggregate score
Example: single statistic – overview of units

Prioritize units within aggregate – based on global score
Example: single statistic – individual unit

Prioritize variables within unit – based on local scores
Example: across statistics – publication aggregates

Prioritize statistics with large inconsistencies between them, for a certain aggregate and mutual variable.
### Example: across statistics – overview of units

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.