What comes next?

Data Provider Forum

Matthias Mertens
Coordinator of the Scientific Team

Virtual Conference
24th April, 2020
Current situation with the code and data

- 7th vintage has improved a lot.
- Rich set of indicators; code became very large
- Running time gets extreme in some countries.
- Burden for some data providers
- For the 8th vintage: Make code more efficient, reduce running time.
- However, we also aim for implementing new variables (and drop existing)
- Who can we do this?
Rich data and speed of computations trade-off

- Compute many interesting things
- Computation speed

Trade-off
Rich data and speed of computations trade-off

Compute many interesting things

Ethics: “Orthogonal positioning” to solve trade-off

IO: Hicks-Neutral technological change

Computation speed
How do we innovate?

- **Core**: Allow to freely adjust which variables are computed in which dimensions. Currently: Each variable for each JD.
- **Expected reduction**: 20%-30%
- **Reduce overall variables and JDs**: To make qualified judgement, we need YOU and the USERS! Which variables are important to you and the users? Feedback is always welcomed.
- **Module 7 takes most time**: We will make the code start form its last break point if it broke.
Data quality

Key Question:
How can we ensure and further advance high quality standards?
Data quality

- Current disclosure routine based on **sample statistics** and done for all variables
- For some variables this is meaningless, e.g. TFP, JDR, JCR, LS
- Here the routine creates too many missing values as mean*number of observations does not give you a total, nor there is a meaningful total of these variables.
- We plan to adjust the dominance criteria to do this only for raw data input and value-added (i.e. revenue, capital, employment,..)
Data quality

- CompNet only reports weighted statistics
- Hence, testing for statistical dominance should be based on **weighted sums**.
- This will greatly increase the amount of data points we can store.
- Addresses problem of “many missing values” in some countries.
- As far as we know: It is a standard that statistical offices test for dominance based on the **relevant population**.
First question of the discussion:
Do you think this is feasible (make the dominance routine based on the relevant population statistics if we report population figures)?

What about running the dominance routine only for the raw data input?

What about minimum number of observations. Would this be feasible to? Theoretically justified if we **do not save** number of firms in sample. What do you think?

Other questions, ideas, suggestions (General discussion)?
Thanks for your attention. Questions and comments are welcome!