Cappadocia 7th International Congress on Scientific Research / August 2-4, 2025 by IKSAD Institute

**Operationalized measurement process for model-quality-evaluation**

**in case of data assets based on questionnaires**

Dr. László Pitlik (<https://orcid.org/0000-0001-5819-0319>),

László Pitlik (Jr.) (<https://orcid.org/0000-0002-8058-9577>),

Mátyás Pitlik (<https://orcid.org/0000-0002-1991-3008>),

e-Mails: pitlik@my-x.hu, ptlklszl@my-x.hu, ppk@my-x.hu

Kodolányi University and MY-X research team Hungary

*Keywords: consistency, automation, questionnaire, randomized data, benchmarking*

**Abstract**

History: The quality of data assets can be massively different: from the randomized data to mechanistic relationships between observed phenomena. In general, questionnaires deliver fewer valuable data than classic measurements. Therefore, it is relevant to be capable of estimating data quality of data assets coming from questionnaires. The previous experiments defined operationalized processes for model-quality-evaluations in order to derive suspicious persons (objects) and/or questions based on the difference between estimation and fact concerning each questionnaire-position.

Background and benchmarks: Here and now, a questionnaire (2011-2022) concerning human beings (20.000+), especially their life-parameters (e.g. sleeping-quality, sport-activity, smoking, etc. – 4 groups as X: 2 binary, 1 unit with three options, 1 unit with 4 options and one attribute as Y – with 5 options: see self-evaluation of the own life quality) are focused in order to construct a model series being capable of highlighting differences between the quality of randomized, real, and ideal data assets, where the randomized and the ideal data assets are the benchmarks and the quality of the real (questionnaire-based) data should be measured on the scale defined by the two extreme alternatives (randomized and ideal data assets).

Highlighted details: The randomized data positions follow the internal relationships of the real data (e.g. binary X-case: ratio of smoking in percent = 100% - ratio of no-smoking / case of 5 Y-options: ration of life quality on level 5 + on level 4 + on level 3 + on level 2 + on level 1 = 100%). The OAM (object attribute matrix) for generation of production functions had 11 objects (years – 2014 is missing) and 11 X-attributes + 5 Y-attributes. A model (Y=f(Xi)) is good, if the Y-values (ratios) in each level of 1-2-3-4-5 can be derived from the constant (ranked) Xi-matrix in a parallel way based on a direction vector for the attributes being defined by experts (by the authors of the questionnaires: e.g. the higher ratio on no-smoking-option leads to higher ration of better life quality levels). The models for each Y-level deliver not only differences between facts and estimations, but also importance of the 11 attributes in each model and/or correlation values between Xi and Yj. Predefined rules for model-evaluation: The ideal data assets should have correlation values exactly as 1.00, the randomized data assets should have lower correlation values as real data assets. The volume of estimation errors should be higher in case of randomized data assets than in case of real data assets. The number of antagonistic objects should be higher in case of randomized data assets than in case of real data assets, etc. Result: the questionnaire-based data asset is better than the randomized one and the real data asset produced no suspicious modelling details – based on a totally algorithmizable/operationalizable process!

Future aspects: The observed data asset through the questionnaires can be seen in this first project layer as a group of relatively good interpretable X-phenomena (c.f. ever smoking yes/no, etc.). It means that these data could have been collected based on classic measurements. On the other hand, there are existing such attributes, where the answers (X-variables) deliver intuitive feelings (like already in case of the Y-variables). It is to expect, that the less measurable X-data will lead to less qualitative modelling results. Therefore, the own “data-quality-meter” should deliver lower index values based on the trivial anti-discriminative principle: Can each data asset have the same quality index?