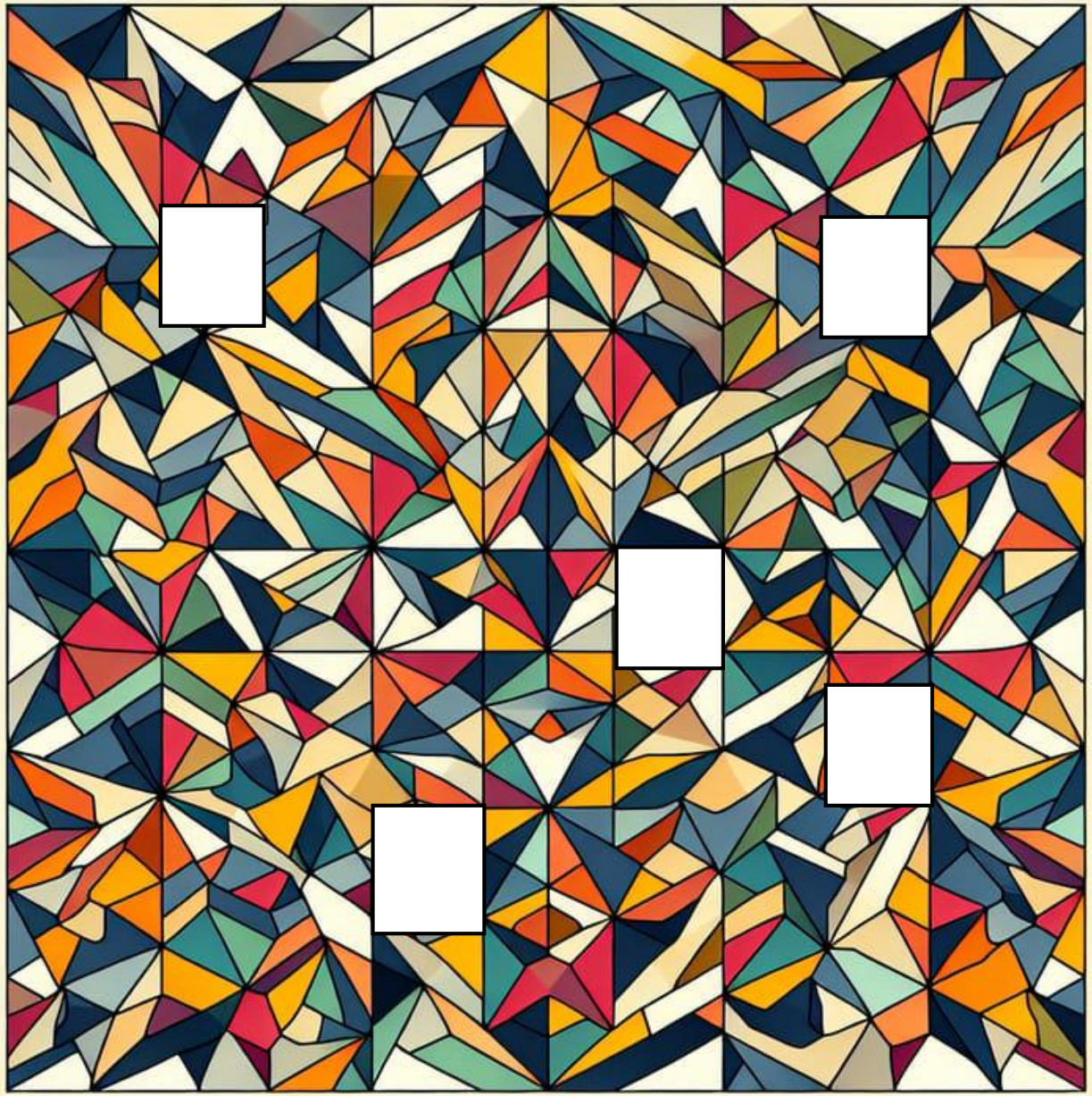


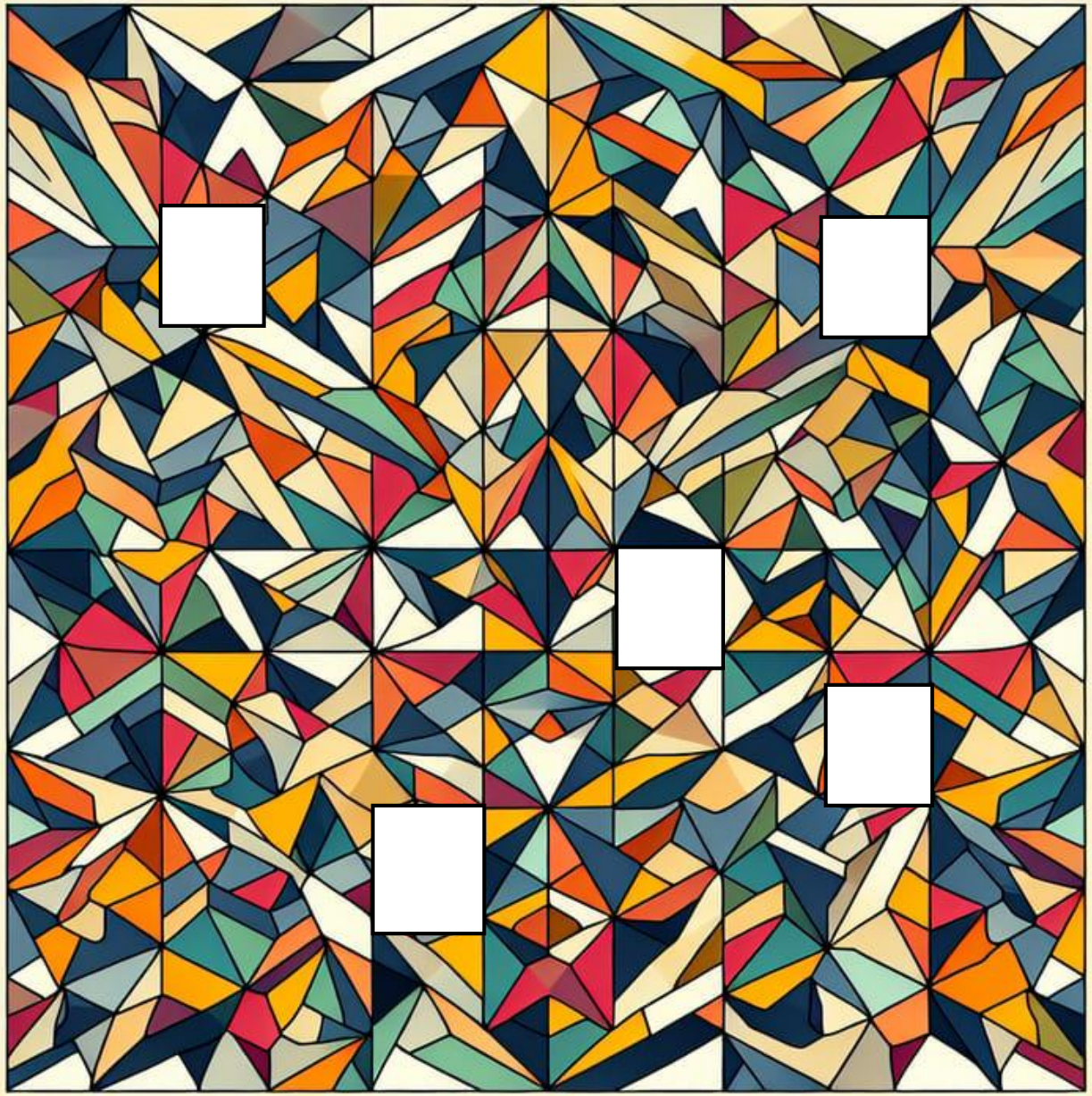
Data imputation methods for InterCriteria Analysis: Whether, What and How?

(research in progress)

Vassia Atanassova
Peter Vassilev

Dept. of Bioinformatics and Biomedical Modelling,
Institute of Biophysics and Biomedical Engineering,
Bulgarian Academy of Sciences

Int. Workshop on Intuitionistic Fuzzy Sets • 13 Dec. 2024



Data imputation methods for InterCriteria Analysis: Whether, What and How?

(research in progress)

In clinical research, nothing can be said to be certain, except for measurement error and missing data.

—Unknown author

Int. Workshop on Intuitionistic Fuzzy Sets • 13 Dec. 2024

InterCriteria Analysis

- Multicriteria multiobject decision support method as of 2014
  - Aimed at the detection of the levels of correlation between any pair of criteria on the basis of the evaluations of objects against these criteria (e.g., criteria reduction)
  - Based on intuitionistic fuzzy sets
  - Exhaustive pairwise comparisons between the evaluations of all objects against all criteria and maintenance of two counters for the cases when the relations (< / > / =) are the same of different:
    - Normalized membership counter for all likewise comparisons (</< or >/>)
    - Normalized non-membership counter for all differences (</>, >/<)
    - The presence of at least one = leads to incrementing the complementary to 1 Uncertainty counter

InterCriteria Analysis

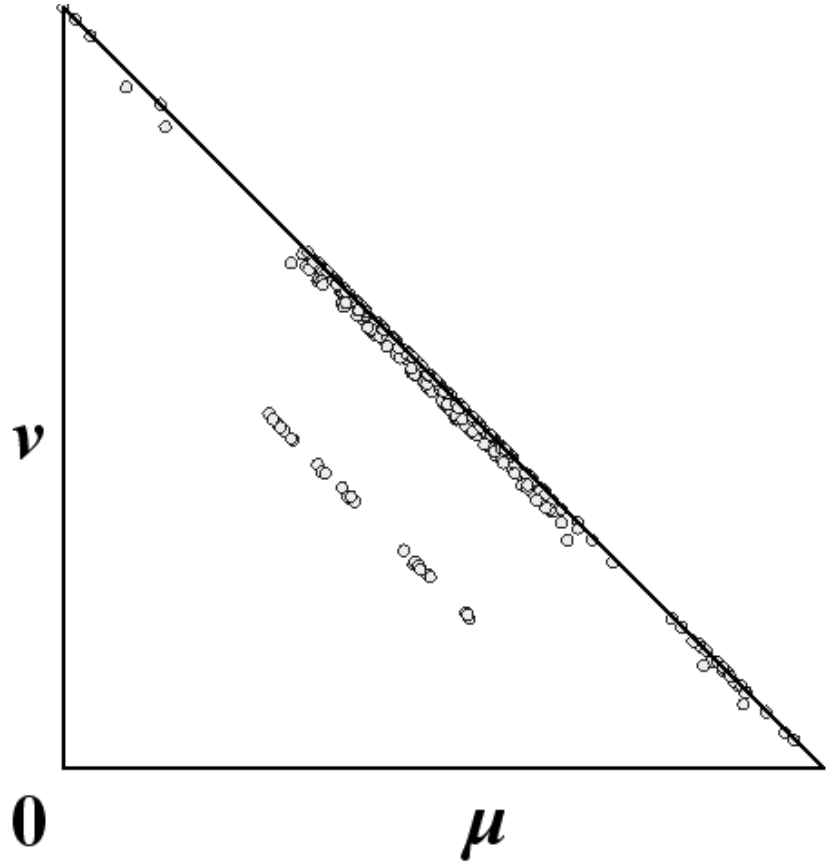
- Input:
  - $m \times n$ table of numerical data of m objects evaluated by n criteria
  - Minimal requirements: at least 3 objects, at least 3 criteria
  - (Problem-specific) threshold values $\alpha, \beta \in [0,1]$
- Output:
  - Two $n \times n$ tables containing the membership and non-membership parts (numbers in $[0,1]$) of the intuitionistic fuzzy evaluations of the intercriteria correlations between each pair of the n criteria
  - The statistical term “correlation” is here interpreted in ICA as positive consonance / negative consonance / dissonance, depending on how the detected membership/non-membership of a pair of criteria stays against the predefined thresholds α / β

InterCriteria Analysis

- An illustrative example of the application of ICA over a dataset of 72 objects by 28 criteria

μ	0 (-)	0 (+)	A1 (-)	A1 (+)	A2 (-)	A2 (+)	B (-)	B (+)	A1B (-)	A1B(+)	A2B (-)	A2B (+)	0	A	A1	A2	B	AB	A1B	A2B	A(-)	A(+)	AB(-)	AB(+)	A1/A	A2/A	A1B/AB	A2B/AB
0 (-)	1.00	0.46	0.41	0.52	0.52	0.45	0.39	0.46	0.51	0.46	0.29	0.53	0.57	0.47	0.49	0.47	0.45	0.50	0.49	0.52	0.40	0.50	0.52	0.48	0.51	0.49	0.46	0.52
0 (+)	0.46	1.00	0.36	0.39	0.50	0.35	0.47	0.35	0.42	0.35	0.30	0.48	0.89	0.34	0.39	0.41	0.36	0.34	0.33	0.47	0.38	0.32	0.43	0.37	0.57	0.43	0.48	0.51
A1 (-)	0.41	0.36	1.00	0.50	0.37	0.55	0.52	0.53	0.51	0.52	0.38	0.49	0.34	0.61	0.63	0.48	0.54	0.54	0.53	0.49	0.84	0.50	0.50	0.52	0.53	0.46	0.50	0.48
A1 (+)	0.52	0.39	0.50	1.00	0.52	0.42	0.41	0.36	0.43	0.47	0.46	0.45	0.39	0.83	0.87	0.38	0.34	0.43	0.45	0.47	0.46	0.85	0.45	0.43	0.64	0.35	0.52	0.46
A2 (-)	0.52	0.50	0.37	0.52	1.00	0.51	0.57	0.37	0.60	0.50	0.46	0.32	0.51	0.51	0.48	0.62	0.41	0.49	0.53	0.33	0.53	0.54	0.62	0.41	0.34	0.64	0.66	0.30
A2 (+)	0.45	0.35	0.55	0.42	0.51	1.00	0.52	0.58	0.58	0.47	0.48	0.42	0.34	0.55	0.42	0.90	0.59	0.49	0.51	0.44	0.58	0.57	0.59	0.43	0.13	0.87	0.54	0.44
B (-)	0.39	0.47	0.52	0.41	0.57	0.52	1.00	0.47	0.55	0.51	0.33	0.44	0.45	0.44	0.39	0.55	0.59	0.54	0.56	0.41	0.62	0.41	0.53	0.49	0.42	0.58	0.59	0.39
B (+)	0.46	0.35	0.53	0.36	0.37	0.58	0.47	1.00	0.40	0.53	0.34	0.61	0.34	0.41	0.37	0.53	0.88	0.53	0.49	0.61	0.50	0.40	0.40	0.57	0.48	0.52	0.37	0.61
A1B (-)	0.51	0.42	0.51	0.43	0.60	0.58	0.55	0.40	1.00	0.56	0.48	0.33	0.42	0.47	0.45	0.62	0.43	0.69	0.72	0.37	0.59	0.50	0.95	0.54	0.37	0.63	0.66	0.32
A1B(+)	0.46	0.35	0.52	0.47	0.50	0.47	0.51	0.53	0.56	1.00	0.30	0.55	0.39	0.47	0.50	0.46	0.52	0.86	0.84	0.50	0.51	0.49	0.52	0.89	0.53	0.47	0.56	0.42
A2B (-)	0.29	0.30	0.38	0.46	0.46	0.48	0.33	0.34	0.48	0.30	1.00	0.28	0.27	0.53	0.47	0.47	0.33	0.33	0.34	0.38	0.37	0.53	0.53	0.28	0.29	0.45	0.37	0.38
A2B (+)	0.53	0.48	0.49	0.45	0.32	0.42	0.44	0.61	0.33	0.55	0.28	1.00	0.52	0.38	0.45	0.37	0.59	0.53	0.46	0.90	0.40	0.40	0.32	0.66	0.63	0.36	0.14	0.84
0	0.57	0.89	0.34	0.39	0.51	0.34	0.45	0.34	0.42	0.39	0.27	0.52	1.00	0.33	0.37	0.41	0.34	0.36	0.36	0.50	0.34	0.33	0.42	0.40	0.56	0.44	0.46	0.52
A	0.47	0.34	0.61	0.83	0.51	0.55	0.44	0.41	0.47	0.47	0.53	0.38	0.33	1.00	0.87	0.49	0.40	0.42	0.45	0.41	0.58	0.88	0.48	0.42	0.54	0.46	0.54	0.44
A1	0.49	0.39	0.63	0.87	0.48	0.42	0.39	0.37	0.45	0.50	0.47	0.45	0.37	0.87	1.00	0.36	0.35	0.44	0.45	0.48	0.57	0.81	0.46	0.46	0.68	0.32	0.50	0.48
A2	0.47	0.41	0.48	0.38	0.62	0.90	0.55	0.53	0.62	0.46	0.47	0.37	0.41	0.49	0.36	1.00	0.57	0.49	0.53	0.39	0.58	0.52	0.63	0.41	0.04	0.96	0.59	0.39
B	0.45	0.36	0.54	0.34	0.41	0.59	0.59	0.88	0.43	0.52	0.33	0.59	0.34	0.40	0.35	0.57	1.00	0.54	0.50	0.57	0.55	0.40	0.42	0.54	0.44	0.56	0.41	0.58
AB	0.50	0.34	0.54	0.43	0.49	0.49	0.54	0.53	0.69	0.86	0.33	0.53	0.36	0.42	0.44	0.49	0.54	1.00	0.92	0.52	0.56	0.45	0.66	0.84	0.49	0.51	0.57	0.42
A1B	0.49	0.33	0.53	0.45	0.53	0.51	0.56	0.49	0.72	0.84	0.34	0.46	0.36	0.45	0.45	0.53	0.50	0.92	1.00	0.44	0.55	0.48	0.68	0.80	0.45	0.55	0.64	0.34
A2B	0.52	0.47	0.49	0.47	0.33	0.44	0.41	0.61	0.37	0.50	0.38	0.90	0.50	0.41	0.48	0.39	0.57	0.52	0.44	1.00	0.39	0.43	0.39	0.61	0.62	0.38	0.08	0.89
A(-)	0.40	0.38	0.84	0.46	0.53	0.58	0.62	0.50	0.59	0.51	0.37	0.40	0.34	0.58	0.57	0.58	0.55	0.56	0.55	0.39	1.00	0.46	0.57	0.48	0.42	0.58	0.58	0.40
A(+)	0.50	0.32	0.50	0.85	0.54	0.57	0.41	0.40	0.50	0.49	0.53	0.40	0.33	0.88	0.81	0.52	0.40	0.45	0.48	0.43	0.46	1.00	0.52	0.43	0.50	0.49	0.55	0.43
AB(-)	0.52	0.43	0.50	0.45	0.62	0.59	0.53	0.40	0.95	0.52	0.53	0.32	0.42	0.48	0.46	0.63	0.42	0.66	0.68	0.39	0.57	0.52	1.00	0.50	0.36	0.64	0.64	0.34
AB(+)	0.48	0.37	0.52	0.43	0.41	0.43	0.49	0.57	0.54	0.89	0.28	0.66	0.40	0.42	0.46	0.41	0.54	0.84	0.80	0.61	0.48	0.43	0.50	1.00	0.58	0.42	0.46	0.52
A1/A	0.51	0.57	0.53	0.64	0.34	0.13	0.42	0.48	0.37	0.53	0.29	0.63	0.56	0.54	0.68	0.04	0.44	0.49	0.45	0.62	0.42	0.50	0.36	0.58	1.00	0.00	0.37	0.61
A2/A	0.49	0.43	0.46	0.35	0.64	0.87	0.58	0.52	0.63	0.47	0.45	0.36	0.44	0.46	0.32	0.96	0.56	0.51	0.55	0.38	0.58	0.49	0.64	0.42	0.00	1.00	0.61	0.37
A1B/AB	0.46	0.48	0.50	0.52	0.66	0.54	0.59	0.37	0.66	0.56	0.37	0.14	0.46	0.54	0.50	0.59	0.41	0.57	0.64	0.08	0.58	0.55	0.64	0.46	0.37	0.61	1.00	0.02
A2B/AB	0.52	0.51	0.48	0.46	0.30	0.44	0.39	0.61	0.32	0.42	0.38	0.84	0.52	0.44	0.48	0.39	0.58	0.42	0.34	0.89	0.40	0.43	0.34	0.52	0.61	0.37	0.02	1.00

ν	0 (-)	0 (+)	A1 (-)	A1 (+)	A2 (-)	A2 (+)	B (-)	B (+)	A1B (-)	A1B(+)	A2B (-)	A2B (+)	0.00	A	A1	A2	B	AB	A1B	A2B	A(-)	A(+)	AB(-)	AB(+)	A1/A	A2/A	A1B/AB	A2B/AB
0 (-)	0.00	0.53	0.58	0.47	0.46	0.54	0.60	0.53	0.48	0.53	0.44	0.46	0.42	0.53	0.50	0.52	0.54	0.50	0.50	0.47	0.59	0.49	0.47	0.51	0.49	0.51	0.52	0.46
0 (+)	0.53	0.00	0.64	0.60	0.48	0.65	0.53	0.64	0.58	0.65	0.43	0.52	0.11	0.66	0.61	0.58	0.64	0.66	0.67	0.53	0.62	0.67	0.57	0.63	0.43	0.57	0.51	0.48
A1 (-)	0.58	0.64	0.00	0.49	0.61	0.45	0.48	0.46	0.48	0.47	0.35	0.50	0.66	0.38	0.36	0.51	0.45	0.46	0.47	0.50	0.16	0.49	0.49	0.47	0.46	0.53	0.48	0.50
A1 (+)	0.47	0.60	0.49	0.00	0.46	0.57	0.58	0.63	0.56	0.52	0.27	0.54	0.60	0.17	0.13	0.61	0.66	0.57	0.54	0.52	0.54	0.14	0.54	0.56	0.35	0.64	0.46	0.52
A2 (-)	0.46	0.48	0.61	0.46	0.00	0.47	0.41	0.61	0.37	0.48	0.27	0.66	0.46	0.47	0.50	0.36	0.57	0.50	0.45	0.64	0.45	0.44	0.35	0.57	0.64	0.34	0.30	0.66
A2 (+)	0.54	0.65	0.45	0.57	0.47	0.00	0.48	0.42	0.42	0.53	0.25	0.57	0.66	0.45	0.58	0.10	0.41	0.51	0.49	0.55	0.42	0.43	0.40	0.57	0.87	0.13	0.44	0.54
B (-)	0.60	0.53	0.48	0.58	0.41	0.48	0.00	0.52	0.44	0.49	0.40	0.55	0.55	0.56	0.61	0.44	0.41	0.46	0.44	0.59	0.38	0.58	0.47	0.51	0.58	0.42	0.39	0.59
B (+)	0.53	0.64	0.46	0.63	0.61	0.42	0.52	0.00	0.59	0.47	0.39	0.38	0.66	0.58	0.62	0.46	0.11	0.46	0.50	0.38	0.49	0.59	0.59	0.42	0.52	0.48	0.61	0.37
A1B (-)	0.48	0.58	0.48	0.56	0.37	0.42	0.44	0.59	0.00	0.43	0.25	0.66	0.57	0.53	0.55	0.37	0.57	0.30	0.27	0.62	0.40	0.50	0.05	0.46	0.63	0.37	0.32	0.66
A1B(+)	0.53	0.65	0.47	0.52	0.48	0.53	0.49	0.47	0.43	0.00	0.43	0.45	0.61	0.53	0.50	0.54	0.48	0.14	0.16	0.49	0.49	0.51	0.47	0.11	0.47	0.53	0.42	0.56
A2B (-)	0.44	0.43	0.35	0.27	0.27	0.25	0.40	0.39	0.25	0.43	0.00	0.45	0.47	0.20	0.27	0.26	0.40	0.40	0.39	0.35	0.36	0.20	0.20	0.46	0.45	0.29	0.37	0.36
A2B (+)	0.46	0.52	0.50	0.54	0.66	0.57	0.55	0.38	0.66	0.45	0.45	0.00	0.48	0.62	0.54	0.63	0.41	0.46	0.53	0.10	0.60	0.59	0.67	0.34	0.36	0.63	0.84	0.14
0	0.42	0.11	0.66	0.60	0.46	0.66	0.55	0.66	0.57	0.61	0.47	0.48	0.00	0.67	0.63	0.58	0.65	0.64	0.64	0.50	0.65	0.67	0.57	0.60	0.44	0.56	0.52	0.46
A	0.53	0.66	0.38	0.17	0.47	0.45	0.56	0.58	0.53	0.53	0.20	0.62	0.67	0.00	0.13	0.50	0.60	0.58	0.55	0.59	0.41	0.12	0.51	0.58	0.46	0.54	0.44	0.54
A1	0.50	0.61	0.36	0.13	0.50	0.58	0.61	0.62	0.55	0.50	0.27	0.54	0.63	0.13	0.00	0.64	0.65	0.56	0.54	0.52	0.42	0.18	0.53	0.54	0.32	0.68	0.48	0.50
A2	0.52	0.58	0.51	0.61	0.36	0.10	0.44	0.46	0.37	0.54	0.26	0.63	0.58	0.50	0.64	0.00	0.43	0.50	0.46	0.60	0.42	0.47	0.36	0.59	0.96	0.04	0.39	0.59
B	0.54	0.64	0.45	0.66	0.57	0.41	0.41	0.11	0.57	0.48	0.40	0.41	0.65	0.60	0.65	0.43	0.00	0.46	0.50	0.42	0.45	0.60	0.57	0.46	0.56	0.44	0.58	0.41
AB	0.50	0.66	0.46	0.57	0.50	0.51	0.46	0.46	0.30	0.14	0.40	0.46	0.64	0.58	0.56	0.50	0.46	0.00	0.07	0.48	0.44	0.55	0.34	0.16	0.51	0.49	0.42	0.57
A1B	0.50	0.67	0.47	0.54	0.45	0.49	0.44	0.50	0.27	0.16	0.39	0.53	0.64	0.55	0.54	0.46	0.50	0.07	0.00	0.55	0.44	0.51	0.32	0.20	0.55	0.45	0.34	0.64
A2B	0.47	0.53	0.50	0.52	0.64	0.55	0.59	0.38	0.62	0.49	0.35	0.10	0.50	0.59	0.52	0.60	0.42	0.48	0.55	0.00	0.60	0.57	0.60	0.39	0.38	0.62	0.89	0.08
A(-)	0.59	0.62	0.16	0.54	0.45	0.42	0.38	0.49	0.40	0.49	0.36	0.60	0.65	0.41	0.42	0.42	0.45	0.44	0.44	0.60	0.00	0.53	0.42	0.52	0.58	0.42	0.40	0.58
A(+)	0.49	0.67	0.49	0.14	0.44	0.43	0.58	0.59	0.50	0.51	0.20	0.59	0.67	0.12	0.18	0.47	0.60	0.55	0.51	0.57	0.53	0.06	0.47	0.56	0.49	0.50	0.43	0.55
AB(-)	0.47	0.57	0.49	0.54	0.35	0.40	0.47	0.59	0.05	0.47	0.20	0.67	0.57	0.51	0.53	0.36	0.57	0.34	0.32	0.60	0.42	0.47	0.00	0.50	0.64	0.36	0.34	0.64
AB(+)	0.51	0.63	0.47	0.56	0.57	0.57	0.51	0.42	0.46	0.11	0.46	0.34	0.60	0.58	0.54	0.59	0.46	0.16	0.20	0.39	0.52	0.56	0.50	0.00	0.42	0.58	0.52	0.46
A1/A	0.49	0.43	0.46	0.35	0.64	0.87	0.58	0.52	0.63	0.47	0.45	0.36	0.44	0.46	0.32	0.96	0.56	0.51	0.55	0.38	0.58	0.49	0.64	0.42	0.00	1.00	0.61	0.37
A2/A	0.51	0.57	0.53	0.64	0.34	0.13	0.42	0.48	0.37	0.53	0.29	0.63	0.56	0.54	0.68	0.04	0.44	0.49	0.45	0.62	0.42	0.50	0.36	0.58	1.00	0.00	0.37	0.61
A1B/AB	0.52	0.51	0.48	0.46	0.30	0.44	0.39	0.61	0.32	0.42	0.37	0.84	0.52	0.44	0.48	0.39	0.58	0.42	0.34	0.89	0.40	0.43	0.34	0.52	0.61	0.37	0.00	0.98
A2B/AB	0.46	0.48	0.50	0.52	0.66	0.54	0.59	0.37	0.66	0.56	0.36	0.14	0.46	0.54	0.50	0.59	0.41	0.57	0.64	0.08	0.58	0.55	0.64	0.46	0.37	0.61	0.98	0.00



InterCriteria Analysis

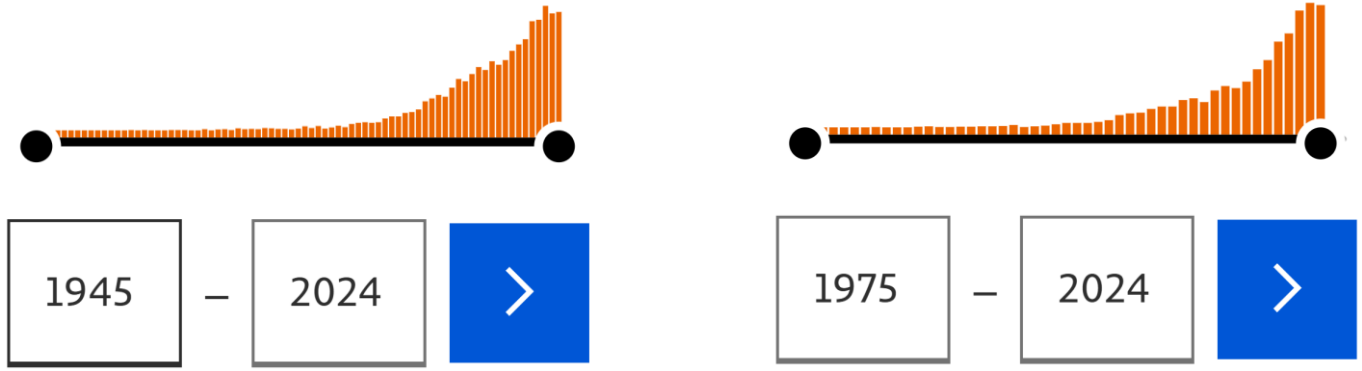
- InterCriteria Analysis has several software implementations
- All of the softwares work with fully populated numerical datasets.
- Successfully tested so far with datasets up to several thousand objects against ~150 criteria

Questions

- How to make ICA work with datasets with some missing data?
- Given that ICA already renders account of uncertainty in its output, then how much missing data would be tolerable in the input, before the results of the ICA get compromised?

Missing data and methods of handling it

- **Missing data** has been a subject of scientific research on its own since mid 20th century with about 5800 papers since 1945, half of them in the last decade, 2016-2024
- **Data imputation** has been a subject of scientific research on its own since mid 1970s, with about 3100 papers since 1975, half of them in the last five years 2020-2024



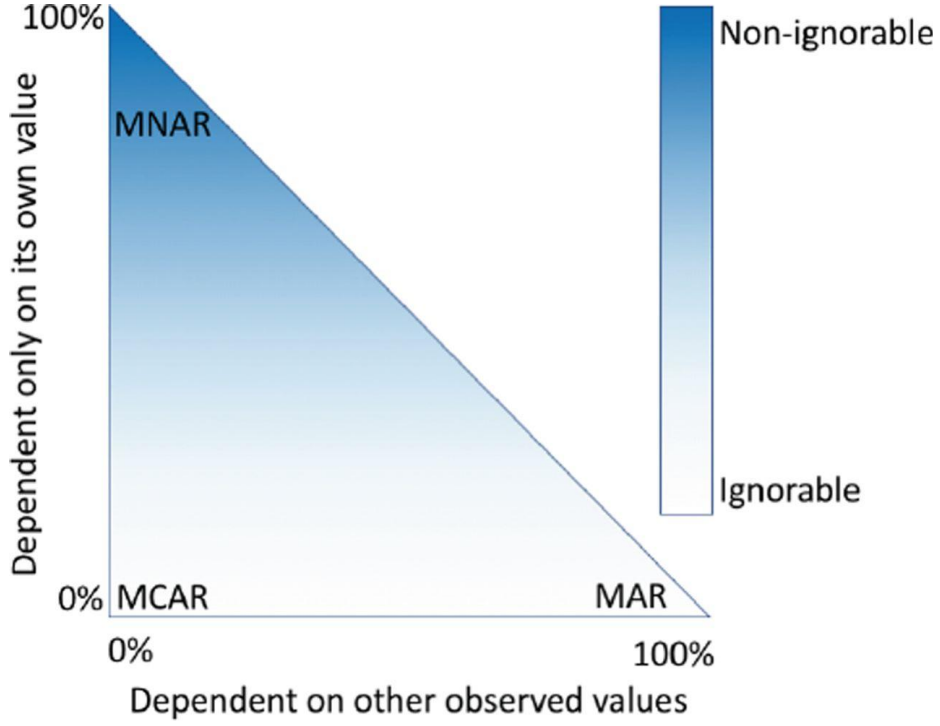
Scopus search by `TITLE ( "missing data" )` and `TITLE ( data AND imputation )`

Missing data and methods of handling it

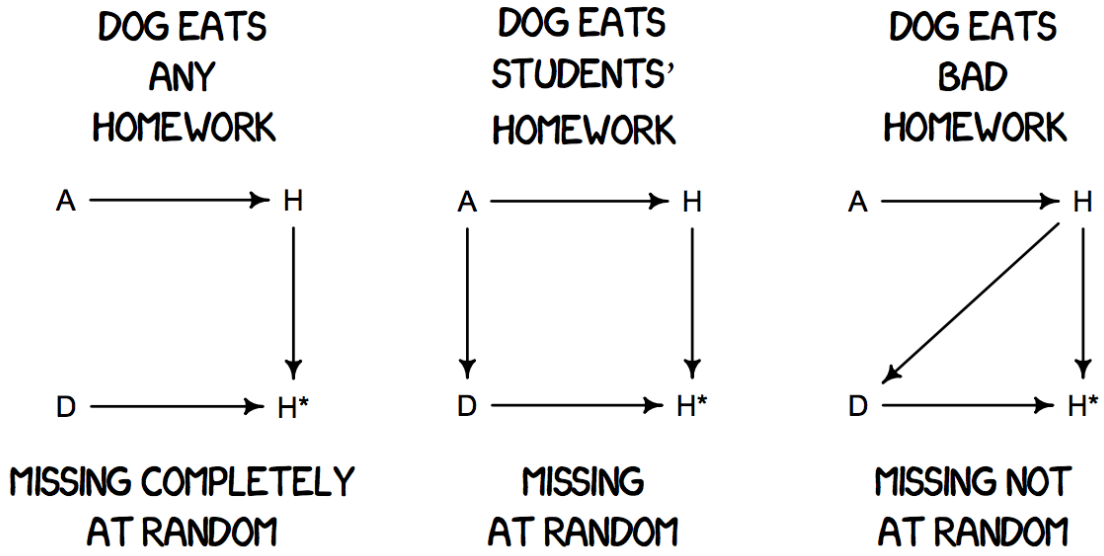
- **Missing data types:**
  - missing completely at random (MCAR) - ignorable
    - the probability of being “missing” is the same for all data, i.e., the causes of the missing data are unrelated to the data itself
    - nothing stops us from assuming that missing values are MCAR if that appears to be a fair thing to do based on the analysis
  - missing at random (MAR) - ignorable
    - the missingness of one feature can be explained by other observed features in the dataset
    - even though the data is missing, its occurrence can still be (somewhat) estimated based on the information available in the dataset
  - missing not at random (MNAR) – non-ignorable
    - the missingness is either attributed to the missing value itself or to feature(s) that we didn’t collect data for
    - there’s not much we can do to address this, except for collecting more data/features; domain expertise becomes extremely important to smartly tackle MNAR and improve the data collection process

Missing data and methods of handling it

- **Missing data types:**



H: Homework
H*: Homework with missing values
A: Attribute of student
D: Dog (missingness mechanism)

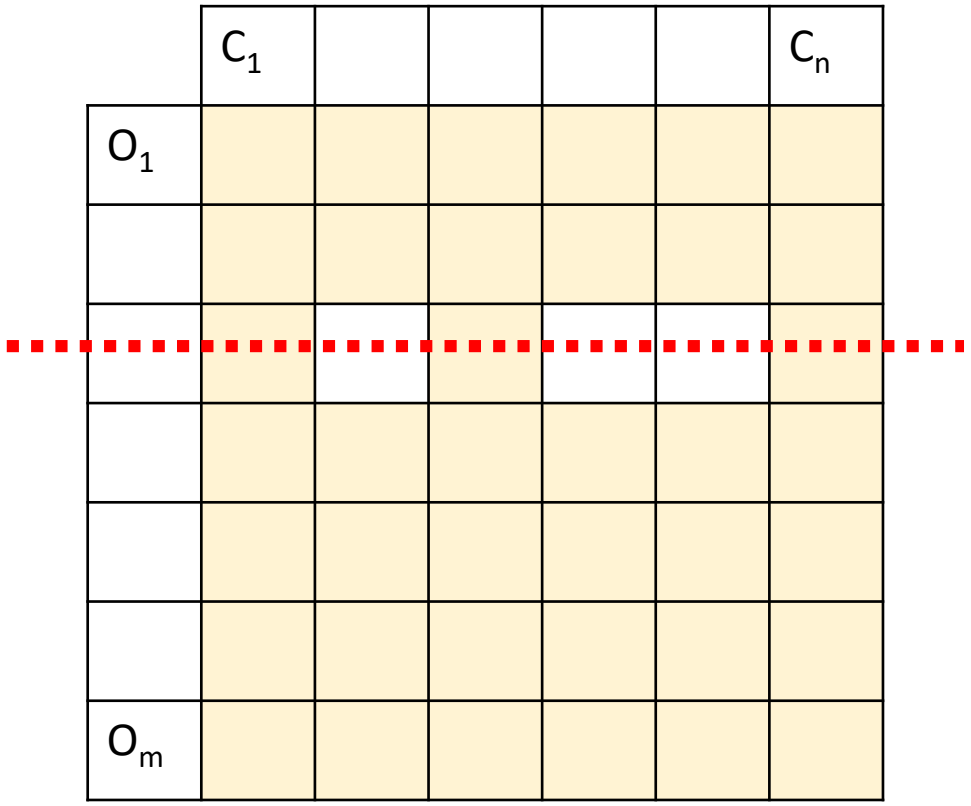


Missing data and methods of handling it

- **Approaches to handling missing data, considered relevant to ICA**
  - Deletion
    - Listwise deletion (objects)
    - Casewise deletion (criteria)
  - Imputation
    - Single imputation, e.g.
      - Mean substitution
      - Last observation carried forward
      - Maximum likelihood estimation
    - Multiple imputation

Deletion of missing data in ICA

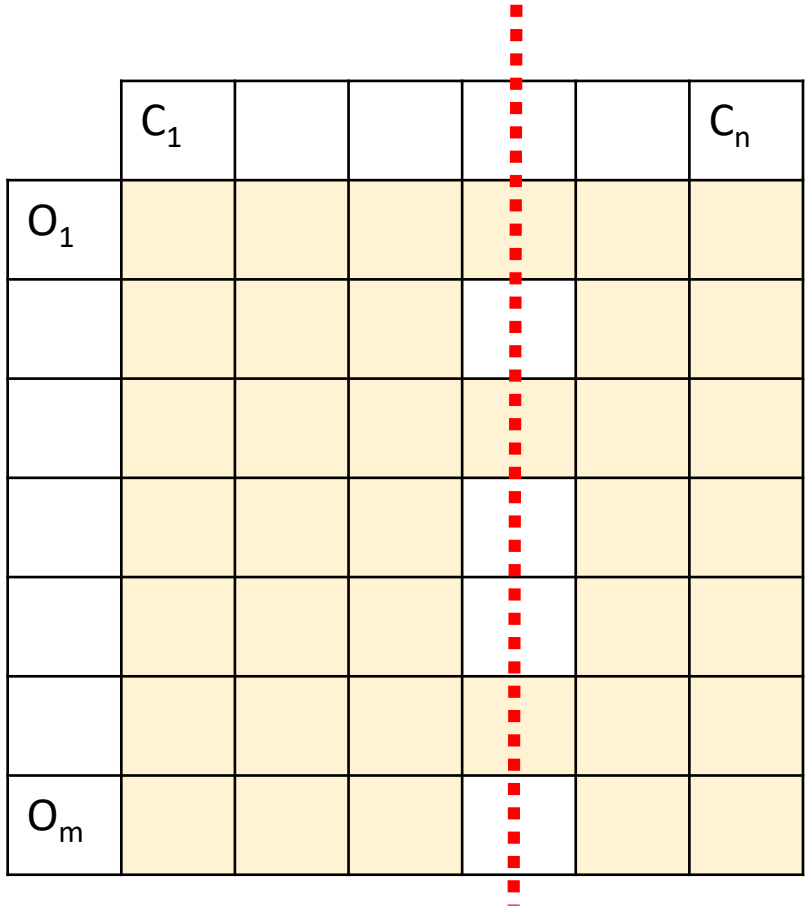
- **Listwise deletion (objects)**



- When the problem is with the object itself: If there is a systemic lack of data for the evaluations of a given object against multiple criteria
- Deletion is fine when we have *reasonably many* objects, i.e. if we remove k objects, the ratio $(m - k) / m$ must remain *reasonably high*
- Also, when every criterion is important and **must** remain in the analysis.

Deletion of missing data in ICA

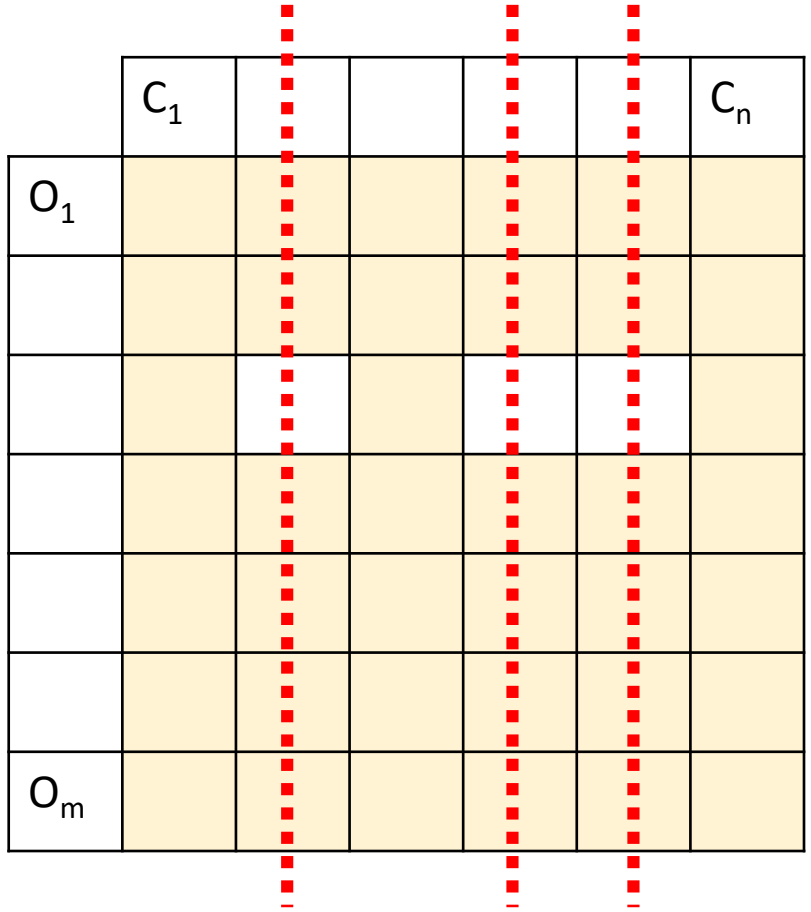
- **Casewise deletion (criteria)**

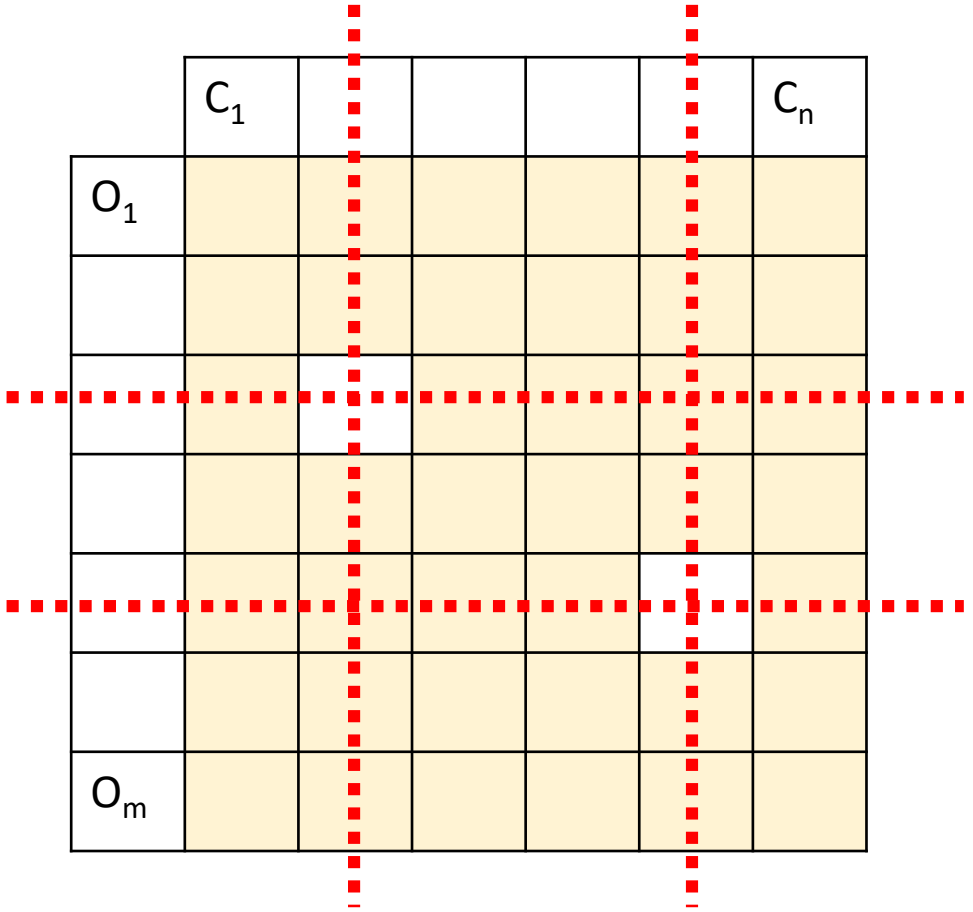


- When the problem is with the criterion itself: If there is a systemic lack of data for the evaluations of multiple objects against the given criterion
- Deletion is fine when we have *reasonably many* criteria, i.e. if we remove l criteria, the ratio $(n - l) / n$ must remain *reasonably high*
- When the criterion to be deleted is **not** one that **must** remain in the analysis.

Deletion of missing data in ICA

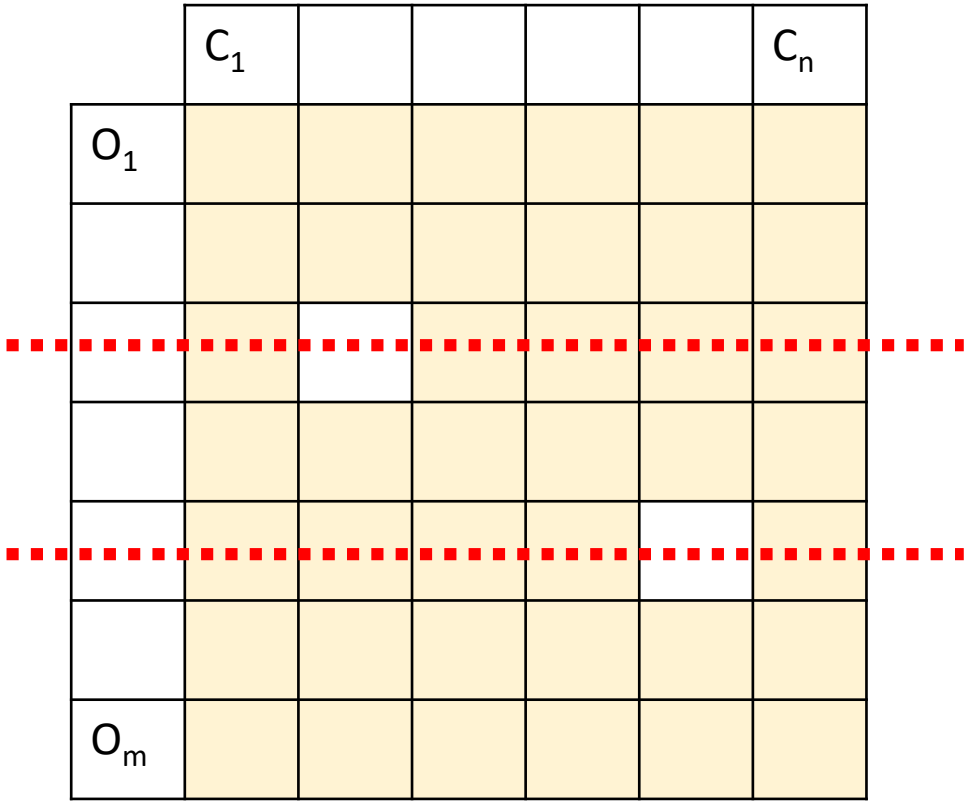
- **Other scenarios?** Possible, but unlikely.

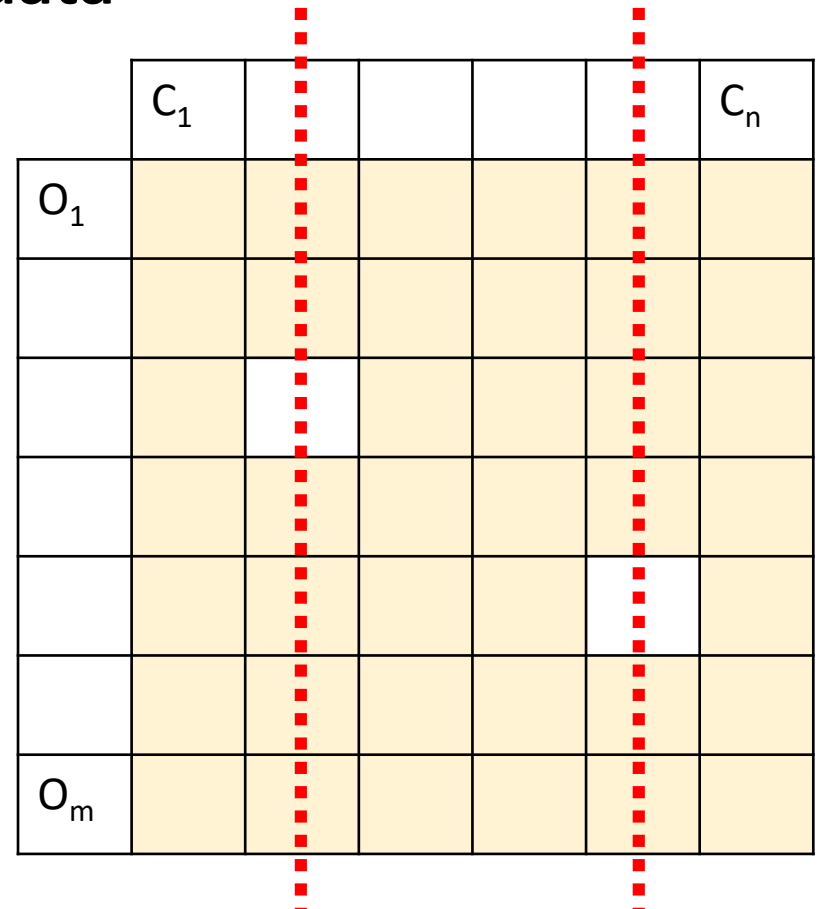




Deletion of missing data in ICA

- **To do: Compare the ICA results in listwise vs casewise deletion with different numbers of missing data**





Missing data imputation in ICA

- Mean substitution
  - Imputation is done once, replacing any missing value with the mean of the non-missing objects' evaluations against that criterion.
  - Pro: Preserves the overall distribution of data.
  - Cons: Can distort relationships between variables and underestimate variability.
  - In ICA this approach would be more appropriate for objects' evaluations featuring less variance (they already exhibit uncertainty in the ICA result, so the increase of uncertainty due to the imputed mean values would not drastically change the result).
  - **To do: Check mean substitution with missing data for criteria exhibiting higher and lower variance and compare**

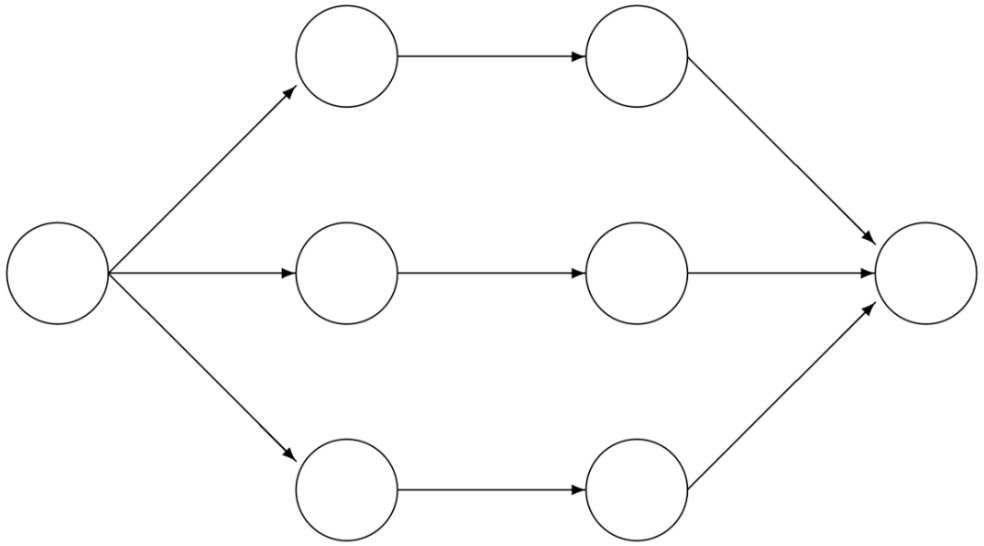
Missing data imputation in ICA

- Last observation carried forward
  - Imputation is done once, replacing any missing value with the last observed value for that object evaluated against that criterion
  - Pro: Simple method.
  - Cons: Relevant only in longitudinal studies; has serious weaknesses and is nowadays less frequently used

  - **To do: Check LOCF only in cases when the object's evaluations against the rest of the criteria exhibits stability over time**

Missing data imputation in ICA

- Multiple imputation
  - Imputation is done a number of times, replacing any missing value with a *plausible* value. ICA is done a number of times, too.

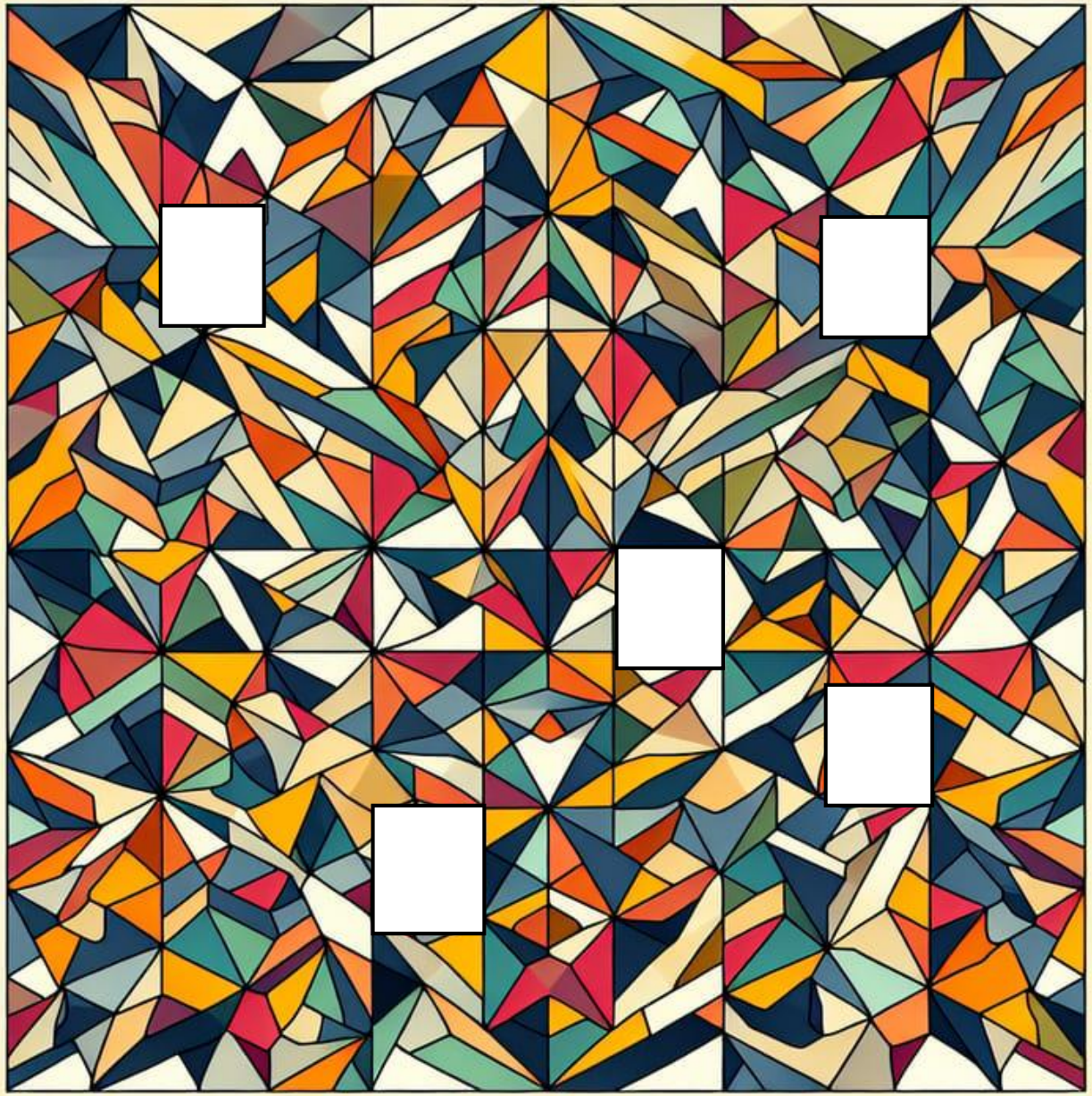


Missing data imputation in ICA

- Multiple imputation
  - Pro: Considered more accurate than single imputation.
  - Cons: More complex and time-consuming; requires careful implementation and interpretation.
  - In ICA this approach would be more appropriate if the results are interpreted as interval-valued intuitionistic fuzzy data.
  - **To do: Check multiple imputation with**
    - **differing proportions of missing data**
    - **the same proportion of missing data and different values**
    - **different threshold values for the ICA results**

Final words

- ICA is an intuitionistic fuzzy sets-based method for decision support which has in its “genetics” the problem of quantifying and handling uncertainty even in full datasets
- Missing data can be another serious challenge in front of certainty and precision, and it stems from:
  - Size of dataset
  - Proportion of missing data
  - Causes of missingness
  - Chosen approach of handling missing data
  - Number of tests conducted with each of the chosen approaches
  - Interpretation of the ICA results against the predefined problem area-specific thresholds for membership and non-membership.



Thank you for your attention!

Vassia Atanassova
Peter Vassilev

vassia.atanassova@gmail.com
peter.vassilev@gmail.com

Acknowledgement to Bulgarian National Science Fund
Grant KP-06-N72/8/2023.