The European Commission's science and knowledge service

Joint Research Centre

European Commission

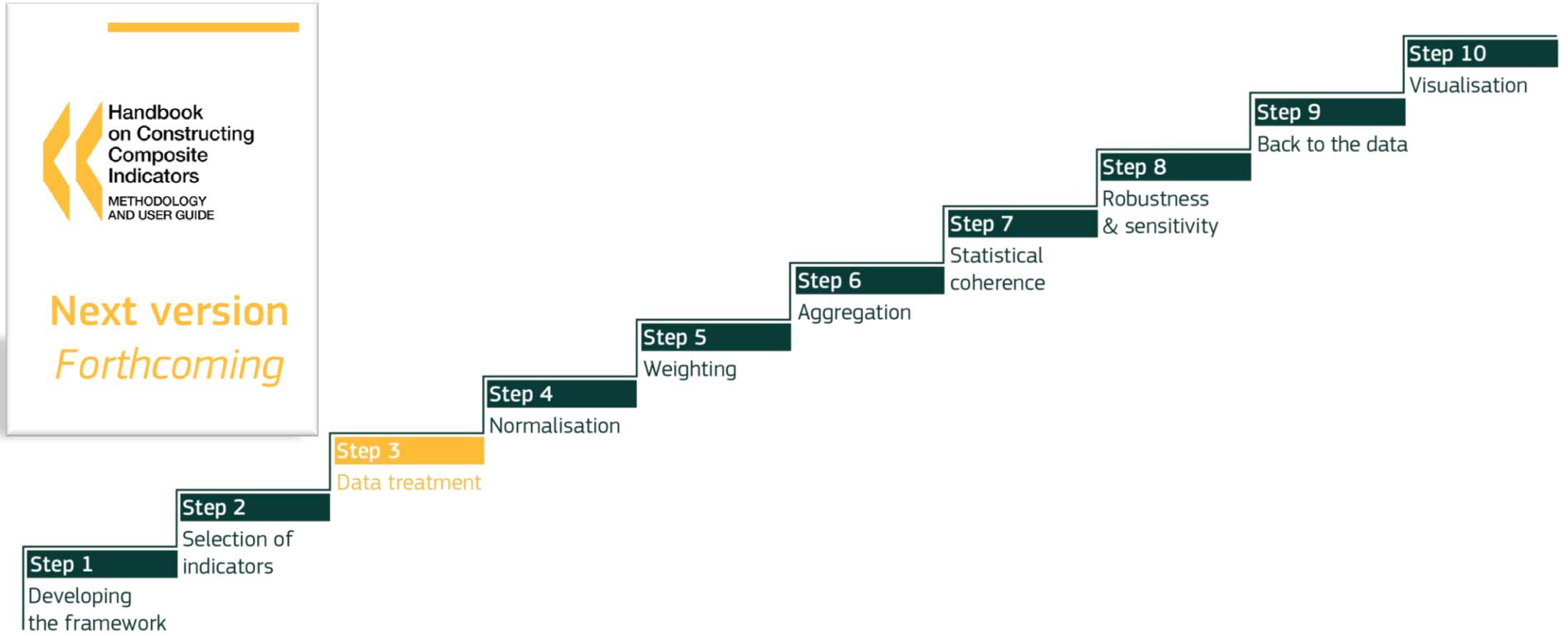
Step 3 (Part II): Missing data

Marcos Domínguez-Torreiro PhD

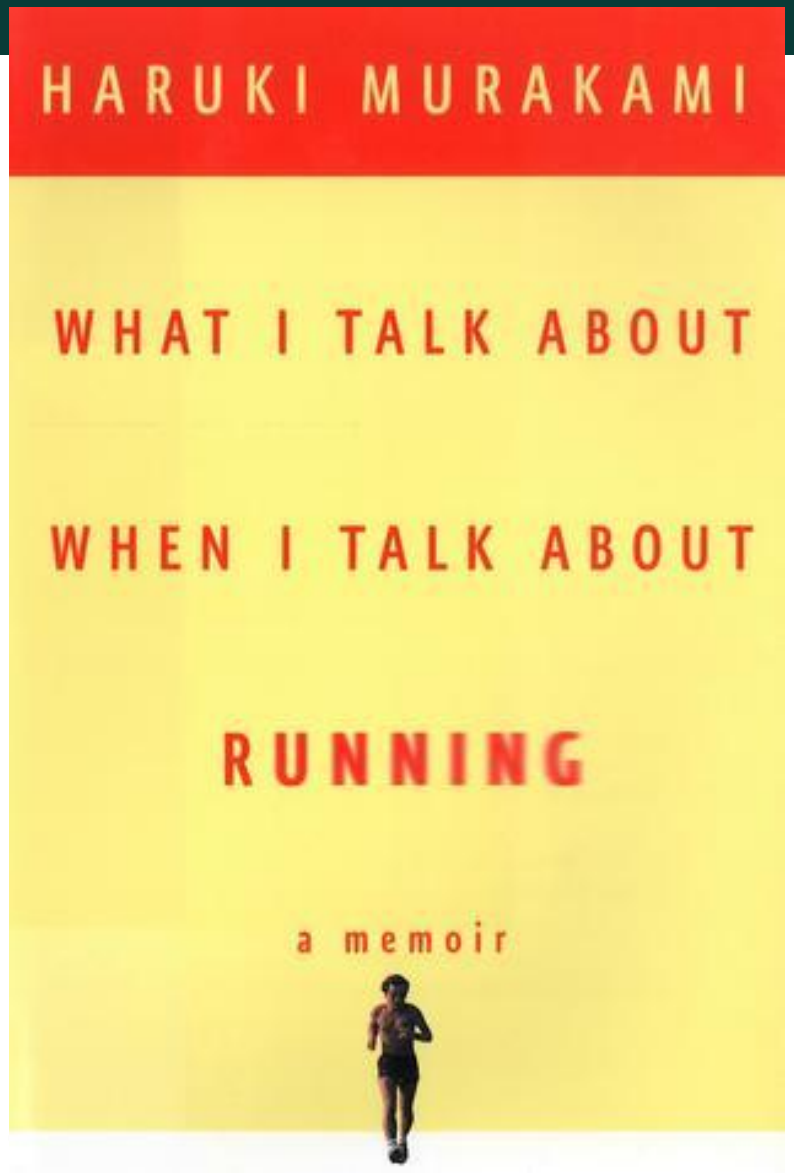
COIN Training 2019

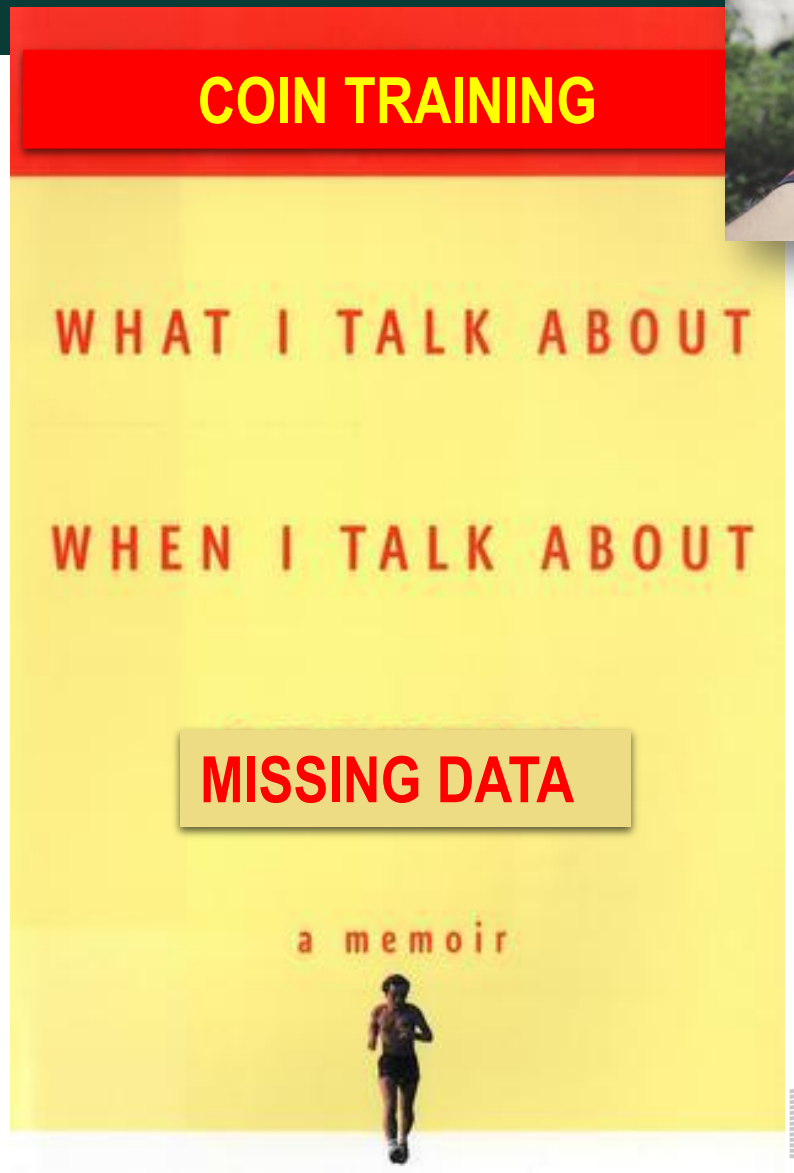
4-6 November, Ispra

Ten steps



Outline





European Commission

Outline

Missing data patterns

- Univariate missing data
- Unit nonresponse
- Item nonresponse

Missing data mechanisms

- MCAR
- MAR
- NMAR

Missing data analysis methods

- Weighting methods
- Deletion methods
- Imputation methods

Definition

"Missing data (or missing values) is defined as the data value that is not stored for a variable in the observation of interest"

Kang, 2013. The prevention and handling of the missing data.

Definition – at a glance

Rows: **units**, cases, observations or subjects

	Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8	Var9	Var10	Missing per country
C1		x			x						2
C2	x		x					x	x	x	5
C3	x	x				x					3
C4											0
C5											0
C6	x		x		x	x		x		x	6
C7											0
C8											0
C9	x			x		x	x		x	x	6
C10											0
C11						x					1
C12											0
C13											0
C14			x		x						2
C15						x					1
C16			x								1
C17											0
C18		x				x					2
C19											0
C20						x					1
Missing per indicator	4	3	4	1	3	7	1	2	2	3	10

Columns: **variables** measured for each unit

Missing data patterns

univariate missing data

- missingness confined to single (outcome) variable: *entries for that variable for some experimental units are missing* [missing values in one of the columns only]
  - ***e.g. experimental data***: no yield recorded for specific crops/plots if seeds did not germinate

unit nonresponse

- entire *survey nonresponse* by subset of individuals [empty rows in the data set]
  - ***e.g. mail survey***

item nonresponse

- *haphazard pattern* of missing values across variables and units in the data matrix [missing values scattered across rows and columns]
  - ***e.g. survey data***, data collection from multiple sources

Example: GTCI 2019, first pillar

Rows: countries

	I1.1.1	I1.1.2	I1.1.3	I1.1.4	I1.1.5	I1.2.1	I1.2.2	I1.2.3	I1.2.4	I1.2.5	I1.2.6	I1.3.1	I1.3.2	I1.3.3	I1.3.4	I1.3.5	I1.3.6	I1.3.7	I1.3.8
unit_01	21.60127	26.8783	58.54554	22.45655	6.756757	8.371315	23.66917	0	nan	11.95929	60.3469	58.29581	0	9.019728	0	4.620972	6.844316	0	nan
unit_02	48.3666	43.28322	73.74435	53.56071	29.72973	66.83812	69.4891	22.3954	3.05867	39.69466	54.37285	36.90446	27.24067	55.55833	47.61876	63.87599	44.87451	21.88142	nan
unit_03	80.42992	71.29144	78.79054	72.94528	75.67568	73.57961	90.51447	82.82709	20.80606	81.80662	84.50295	89.12535	70.85894	67.79722	75.53606	84.76018	78.9236	79.56622	0.895269
unit_04	50.4218	46.92915	68.86849	40.93937	35.13514	46.75757	50.35727	30.64141	11.35908	66.03053	90.65229	88.09653	29.42132	15.1433	53.39819	32.23693	44.02728	34.65555	5.174625
unit_05	44.19753	49.03165	49.52551	54.97095	28.37838	77.25748	79.95713	42.40332	4.67175	61.57761	57.62663	41.15286	38.85673	51.64526	49.86237	54.12834	48.58096	46.25948	nan
unit_06	83.76485	91.93819	84.69283	95.46467	85.13514	86.58878	88.46016	55.83927	41.8366	80.40712	83.91566	88.82568	76.8501	39.20628	95.83055	67.49376	75.90923	65.7643	21.91873
unit_07	81.96521	94.98874	87.76721	83.36774	83.78378	85.96478	85.67345	78.08227	68.93955	85.24173	52.04831	89.48529	97.59145	77.41804	81.47306	67.55246	77.20703	64.42384	51.72729
unit_08	42.62445	39.68875	48.31945	42.05639	14.86486	56.62558	85.7985	59.06759	3.74025	62.84987	49.03897	80.97789	62.97883	57.35152	58.06123	73.55683	64.05617	67.18223	nan
unit_09	14.07142	20.07524	21.75318	27.56527	4.054054	35.14103	30.01072	23.71567	2.338286	5.852417	0	51.04902	20.01337	28.43122	29.15764	24.02746	1.54371	23.47869	nan
unit_10	75.18949	84.03577	74.31619	78.62692	82.43243	85.64292	77.42051	74.95738	56.81448	82.31552	97.70157	88.72667	71.48233	47.88481	87.19858	62.32376	73.72726	67.50337	55.74589
unit_11	32.26696	43.2165	44.70589	37.19947	36.48649	50.62543	37.44194	20.6722	14.37654	14.50382	18.77309	62.11208	8.115095	26.81464	16.14396	16.65127	25.78795	23.91937	nan
unit_12	28.65253	36.99057	37.58512	28.27211	16.21622	67.09783	20.29296	43.85691	nan	17.43003	27.13678	64.60119	25.28294	33.92418	40.70152	45.5555	45.42924	29.06799	nan
unit_13	52.87184	51.70895	73.20888	63.44682	37.83784	63.07038	72.57949	43.77312	16.57781	65.52163	71.26316	92.485	44.98063	18.61908	30.85267	43.56508	48.57635	46.60544	2.340609
unit_14	51.04856	63.09581	44.19637	58.26076	29.72973	71.44983	70.09646	67.06145	1.897638	82.1883	87.68207	92.00453	67.32887	67.43946	63.53982	70.92873	70.43806	nan	nan
unit_15	34.85376	47.72978	56.62508	44.40937	32.43243	57.6773	59.32476	27.60826	4.053746	52.92621	40.48993	49.25926	27.02927	10.86446	20.45752	23.56995	36.85114	16.99086	0.549058
unit_16	37.0723	24.44941	58.35777	26.02802	20.27027	60.88808	35.209	16.81652	3.121807	34.86005	64.84354	80.25007	6.799115	10.49406	24.06592	16.92199	24.39666	12.03045	nan
unit_17	39.56225	46.02418	56.12137	45.27782	28.37838	68.26128	52.51876	48.85191	27.42938	58.14249	84.55652	77.43712	26.53616	19.31561	53.56084	33.3278	53.42522	38.95555	4.005766
unit_18	74.28409	67.76948	91.12248	65.64764	66.21622	56.41531	73.99071	34.76793	nan	73.66412	74.27711	75.38808	54.38461	56.74737	45.2068	45.60114	46.25062	23.1991	nan
unit_19	60.05553	67.31844	89.77119	39.92578	72.97297	56.51303	63.36191	39.37501	nan	30.66158	32.03892	39.65284	61.89551	62.36656	55.06888	59.40538	40.89081	35.82888	nan
unit_20	56.84691	64.66887	87.60658	59.48843	63.51351	57.36542	62.1472	22.73874	11.45861	40.96692	64.86818	77.30235	31.74674	35.4416	57.10097	37.26344	36.9715	30.83934	nan
unit_21	91.48768	94.72543	89.3935	94.39186	90.54054	78.67368	86.90604	69.87254	34.43415	79.51654	78.62573	84.84185	75.33013	67.51885	92.40134	78.60982	76.73897	68.25103	50.85717
unit_22	96.38587	97.68401	91.55197	94.32477	95.94595	80.47952	80.52876	91.86313	73.65405	91.22137	69.87102	88.6376	100	100	96.7612	95.38438	96.14877	89.43152	43.31701
unit_23	67.17821	76.29025	73.31525	80.85748	71.62162	78.70184	73.59771	39.09833	7.625629	65.01272	85.70059	76.89969	33.80345	39.22455	68.37796	62.41142	66.95126	38.41913	0.698615
unit_24	56.76761	46.52411	59.50324	44.46297	33.78378	78.50957	76.86674	65.86412	46.34481	49.61832	53.03064	nan	70.11912	46.17288	57.11401	68.62243	54.02696	59.71479	39.65368
unit_25	28.56637	37.92775	40.99959	39.20415	28.37838	71.61415	48.92819	19.16471	1.714363	28.49873	43.40291	59.12562	11.69958	48.06362	43.09101	30.66564	56.03964	28.47484	nan

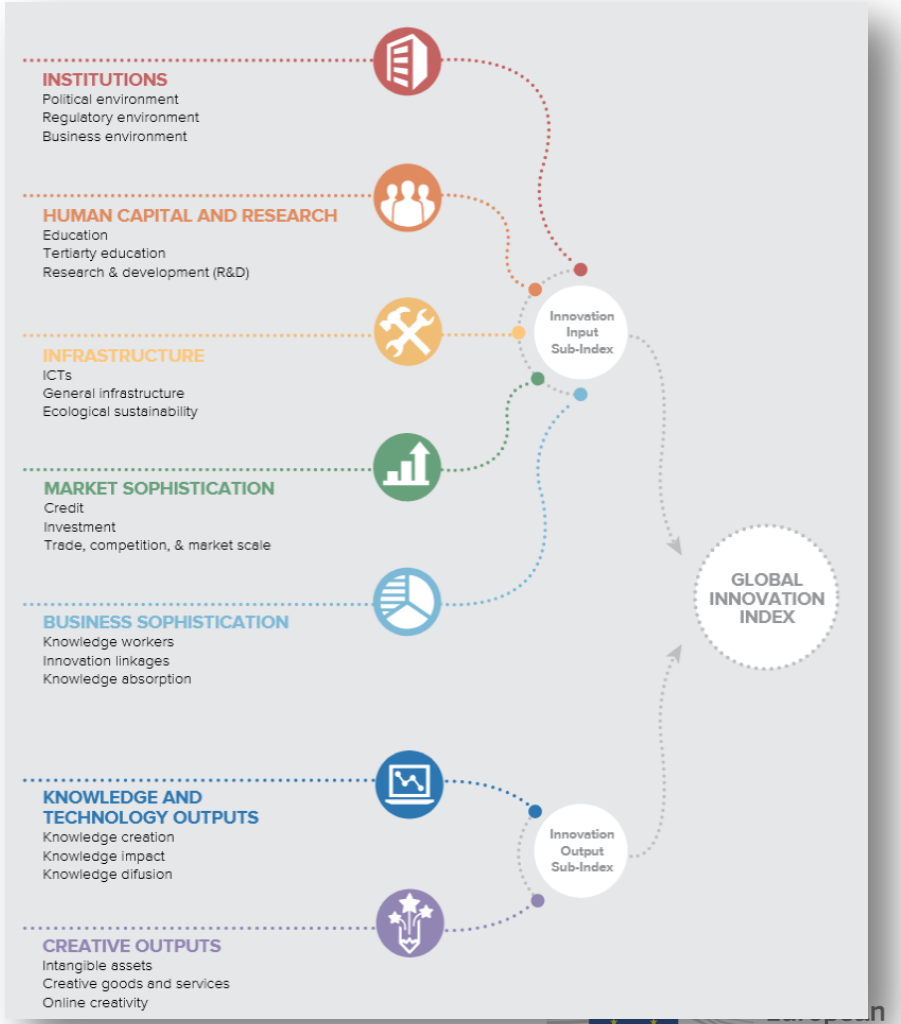
Columns: indicators (normalised values 0-100)

Example: GII 2019

Units (countries) and indicators are selected taking into account pre-defined missing data thresholds:

- @Unit (country)-level:
  - at least 2/3 of indicators available within each of the two sub-indexes (Innovation Input and Innovation Output)
  - scores available for at least two of the three sub-pillars in each pillar

- @Indicator-level: approx. 75 countries with valid cases out of 129…with exact thresholds defined by developers based on indicator importance



Missing data “mechanisms” (= distribution of missingness) – statistical relationships between probability of missingness and missing/observed values in the data set

MCAR – Missing completely at random

- missingness does not depend on the values in the data matrix, missing or observed
  - observed units random subsample of original sample - ***values missing randomly***
    - ***e.g. rolling a die*** *(unrelated to any variable in the data matrix)*

MAR – Missing at random

- missingness depends on observed components and not on the missing components
  - observed units not random sample of original sample - ***values missing systematically***
  - potentially unbalanced data in categories/subpopulations (i.e. contingent emptiness of cells)
    - ***e.g. missing income related to ethnicity and education*** (***fully recorded*** in the data set)

NMAR – Not missing at random

- missingness depends on missing values in the data matrix (either missing values of variable itself or other partially unobserved variables)
  - observed units not random sample of original sample - ***values missing systematically***
    - ***e.g. missing income related to income level***

What does it mean in practice?

- The ***MAR assumption***, although perhaps unrealistic, is generally a ***better approximation to reality than*** the ***MCAR assumption.***

- In general, there is no way to accurately test whether MAR holds in a data set, but…

  … analysis methods appropriate ***under general MAR conditions*** (e.g. imputation methods based on ***likelihood approaches or*** parametric ***MI methods***) ***are still pretty good when dealing with*** missing data that are ***not MAR***.

  (see Schafer and Graham, 2002; Little and Rubin, 2002)

No-imputation methods for handling missing data

weighting procedures

- used in ***sample surveys*** to handle ***unit nonresponse*** (i.e. when all survey items are missing for those in the sample that did not participate)
- partially adjust for nonresponse bias by assigning weights to the respondents
- e.g. randomization inference in surveys, propensity weighting, post-stratification, etc.

complete case analysis / listwise **deletion**

- confine the analysis to the set of cases with no missing values / ***discard incompletely recorded units***
- might be satisfactory with small amounts of missing data
- can lead to serious biases in the analysis if missing data not MCAR, especially when drawing inference for subpopulations

pairwise **deletion**

- ***ignore missing data*** (i.e. for each unit only observed values are considered)
- ***default option*** amongst ***CI practitioners***
- ***"do nothing" == averaging indicators*** (in a given pillar) ***for a given unit***

"shadow imputation"

Deletion methods - comparison

Listwise deletion - discard

Country	Pupil-teacher ratio, secondary	Tertiary enrolment	Graduates in science and engineering
DNK	11.3	81.5	20.4
~~SGP~~	~~14.9~~	~~69.8~~	~~N/A~~
FIN	12.8	87.3	27.9
~~DEU~~	~~12.1~~	~~68.3~~	~~N/A~~
~~IRL~~	~~N/A~~	~~77.6~~	~~23.8~~
KOR	15.6	95.3	31.9
~~ISL~~	~~N/A~~	~~81.3~~	~~15.6~~

Pros: simple; comparability of univariate statistics (same number of cases for all indicators)
Cons: loss of information; loss of precision, and bias when missing data mechanism is not MCAR (complete cases not a random sample of all cases)

vs.

Pairwise deletion - ignore

Country	Pupil-teacher ratio, secondary	Tertiary enrolment	Graduates in science and engineering
DNK	11.3	81.5	20.4
SGP	14.9	69.8	~~N/A~~
FIN	12.8	87.3	27.9
DEU	12.1	68.3	~~N/A~~
IRL	~~N/A~~	77.6	23.8
KOR	15.6	95.3	31.9
ISL	~~N/A~~	81.3	15.6

Pros: simple; retains more data compared to listwise deletion
Cons: it is data treatment even if it goes unnoticed **("shadow imputation")**; might encourage countries not to report bad performances

“Shadow imputation” – behind the scenes

“shadow imputation” == assigning mean value of indicators observed for that unit (row mean)

Note that pillar averages based only on observed values only are identical to pillar averages after imputing row mean values

Country	Pupil-teacher ratio, secondary	Tertiary enrolment	Graduates in science and engineering	Mean
DNK	11.3	81.5	20.4	37.7
SGP	14.9	69.8	~~N/A~~	42.4
FIN	12.8	87.3	27.9	42.7
DEU	12.1	68.3	~~N/A~~	40.2
IRL	~~N/A~~	77.6	23.8	50.7
KOR	15.6	95.3	31.9	47.6
ISL	~~N/A~~	81.3	15.6	48.5

Country	Pupil-teacher ratio, secondary	Tertiary enrolment	Graduates in science and engineering	Mean
DNK	11.3	81.5	20.4	37.7
SGP	14.9	69.8	42.4	42.4
FIN	12.8	87.3	27.9	42.7
DEU	12.1	68.3	40.2	40.2
IRL	50.7	77.6	23.8	50.7
KOR	15.6	95.3	31.9	47.6
ISL	48.5	81.3	15.6	48.5

Imputation methods

implicit modelling

- substitution
- cold deck
- hot deck (e.g. kNN)

explicit modelling

- mean imputation (unconditional or conditional)
- regression imputation
- likelihood based approaches (e.g. EM algorithm)

multiple imputation

- creates *m*-imputed data sets (e.g. Amelia II software)

Implicit modelling

Substitution – at fieldwork stage

Replacing nonresponding units in a field survey with other units not previously selected into the sample (e.g. household in the same block)

Cold & Hot deck (etimology)

Punch cards used in the early days of computing; "donors" for item nonresponse taken from the sample/deck being processed ("hot") vs pre-processed /external data sources ("cold")

Implicit modelling: Nearest Neighbour (kNN) hot deck

Replaces missing values for a nonrespondent (***recipient***) with observed values from a respondent (***donor***) "similar" (based on distance metrics) to the recipient with respect to observed characteristics

Step 1. Compute the distance / similarity between recipient and potential donors

Manhattan (absolute) distance preferred option if high differences shall not be overweighed; alternative metrics: *Euclidean* (square), Mahalanobis, etc.

Country	Expenditure on education	Government expenditure on education per pupil, secondary	School life expectancy
SGP	2.9	16.7	12.8
DEU	4.9	23.7	17.3
IRL	5.3	26.0	19.0
KOR	4.6	23.4	16.6
ISL	7.8	18.3	19.6
LUX	4.1	19.4	13.9
JPN	3.8	25.1	15.4
FRA	5.5	26.8	16.3
HKG	3.3	20.4	N/A

Normalized values

$$d_{ij} = \sum_k \left| {}_k x_i - {}_k x_j \right| \quad \text{Manhattan}$$

Index k goes through all the indicators jointly observed on units i and j

$$d_{ij} = \sqrt{\sum_k \left({}_k x_i - {}_k x_j \right)^2} \quad \text{Euclidean}$$

Country	Distance	
	Euclidean	Manhattan
SGP	3.68	4.02
DEU	3.74	5.02
IRL	6.00	7.70
KOR	3.31	4.37
ISL	4.97	6.55
LUX	1.30	1.83
JPN	4.79	5.27
FRA	6.85	8.72
HKG	0	0

closest country
2nd closest country
3rd closest country

Step 2. The imputed value for the recipient is the observed value on the most similar unit, or the mean value of the *k*-closest

Number of neighbours	Distance type	Imputed value	
1NN	Euclidean	13.9	
1NN	Manhattan	13.9	
2NN	Euclidean	15.3	[= (13.9+16.6)/2]
2NN	Manhattan	13.4	[= (13.9+12.8)/2]
3NN	Euclidean	17.4	[= (13.9+16.6+12.8)/3]
3NN	Manhattan	17.4	[= (13.9+12.8+16.6)/3]
...	...	...	

Pros: uses actual values (easy to communicate); does not impose a structure on relationships between variables.
Cons: might be computational-intensive; might reduce variance, but typically less than mean substitution.

Explicit modelling: Mean (*or median or mode*) imputation

Unconditional - Replaces missing values for an indicator with the arithmetic mean of the observed values for that indicator (mean-column); ***Conditional*** – When units are quite heterogeneous, respondents and nonrespondents are grouped in "similar" classes based on observed variables, and averages are calculated *within* those relatively homogeneous groups of observations

Unconditional mean imputation (by column)

Country	Pupil-teacher ratio, secondary	Tertiary enrolment	Graduates in science and engineering
DNK	11.3	81.5	20.4
SGP	14.9	69.8	N/A
FIN	12.8	87.3	27.9
DEU	12.1	68.3	N/A
IRL	N/A	77.6	23.8
KOR	15.6	95.3	31.9
ISL	N/A	81.3	15.6

Country	Pupil-teacher ratio, secondary	Tertiary enrolment	Graduates in science and engineering
DNK	11.3	81.5	20.4
SGP	14.9	69.8	23.9
FIN	12.8	87.3	27.9
DEU	12.1	68.3	23.9
IRL	13.3	77.6	23.8
KOR	15.6	95.3	31.9
ISL	13.3	81.3	15.6
Mean	**13.3**	**80.2**	**23.9**

Pros: simple, relies on the observed values from the same variable.

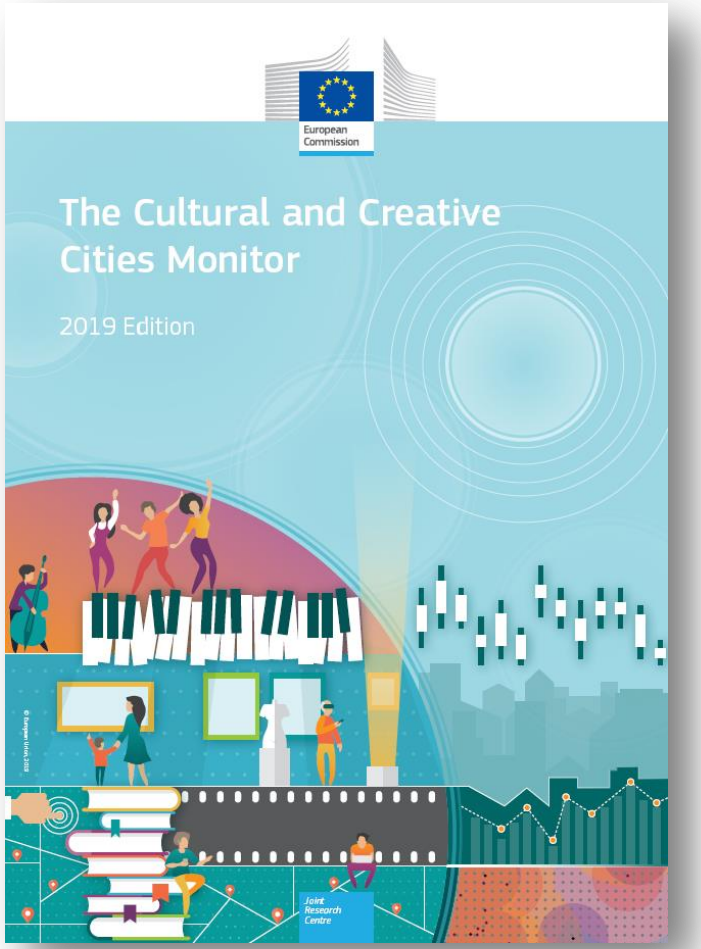
Cons: correlations are affected; variances will be typically under-estimated (as missing values are imputed with 'central values').

Example: The Cultural and Creative Cities Monitor 2019

Missing data filled in using a ***two-step*** (mixed) approach:

- Conditional mean: For missing values in perception-based variables related to foreigners and trust, uses the average value of cities in the same country. Imputed 37% of the 1 064 missing values in the dataset.

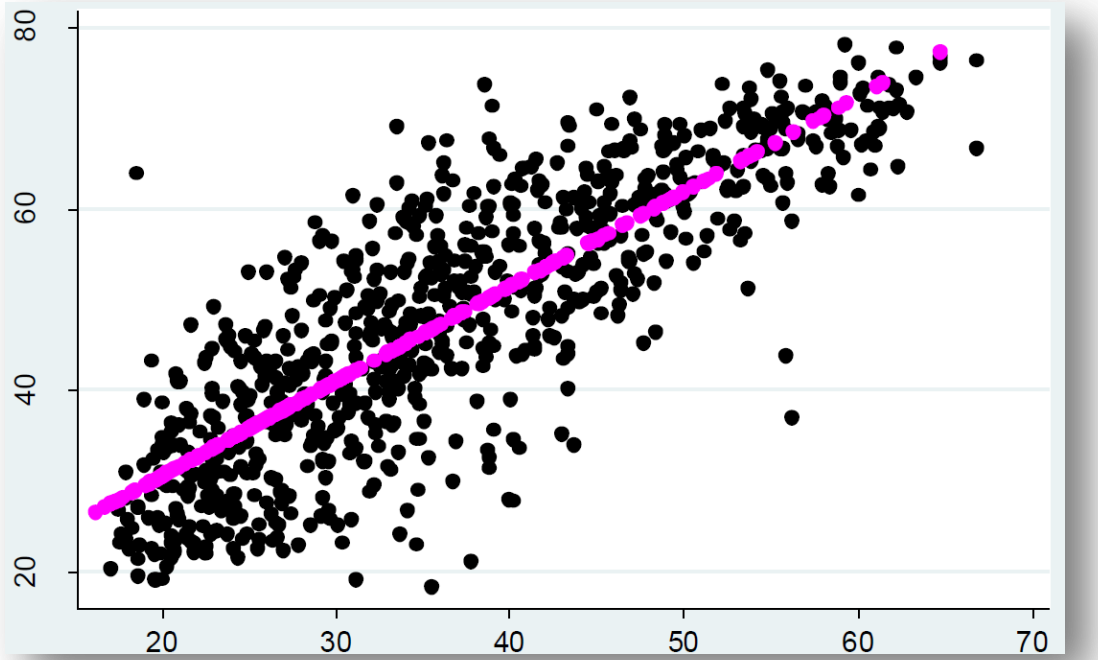
- Hot deck: Remaining missing values imputed using 3NN method, i.e. average value of the 3 cities statistically closest. Donors identified using the *triplet population-GDP-employment rate*. Imputed 63% of the 1 064 missing values in the dataset.



European Commission

Explicit modelling: Regression imputation

The missing variables for a unit are estimated by predicted values from a *multiple linear regression* model on the known variables for that unit.



Any missing values for indicator Y (vertical axis) would have been imputed using the regression model estimates (in purple)

Pros: uses information from observed data.
Cons: imposes a linear structure on relationships between variables; overestimates model fit and correlation estimates; weakens variance.

Explicit modelling – The Expectation Maximisation (EM) algorithm

- Likelihood based approaches: defining a (parametric) model for the observed data and estimating those parameters by ***Maximum Likelihood (ML)***
- ***EM is an iterative procedure*** to compute ML estimates from incomplete data sets (*intuitive idea: filling in missing values and iterating*)

- Each iteration of EM until convergence consists of two-steps:
  - ✓ ***E-step***: calculates conditional expectation of missing data given observed data and current estimated parameters
  - ✓ ***M-step***: ML estimation of parameters as if there were no missing data (i.e. maximizing likelihood of the "expected complete-data")

Pros: works well with **strong correlations (aprox. 0.6)**; might provide unbiased imputed values even if NMAR
Cons: highly **dependent on correlation structure**; difficult to communicate, computational-intensive (but increasingly automatised in statistical software)

Example: GTCI 2019, first pillar

Missing data imputation is usually performed on already normalised variables—for practical purposes. When EM returns out-of-bound values, they are easily detected. Proceed by either "capping" the imputed values (0 – 100) or deleting prior to running kNN only on those missing values.

Rows: countries

unit_01	21.60127	26.8783	58.54554	22.45655	6.756757	8.371315	23.66917	0	-16.691	11.95929	60.3469	58.29581	0	9.019728	0	4.620972	6.844316	7.11E-15	-26.2486
unit_02	48.3666	43.28322	73.74435	53.56071	29.72973	66.83812	69.4891	22.3954	3.05867	39.69466	54.37285	36.90446	27.24067	55.55833	47.61876	63.87599	44.87451	21.88142	-5.73449
unit_03	80.42992	71.29144	78.79054	72.94528	75.67568	73.57961	90.51447	82.82709	20.80606	81.80662	84.50295	89.12535	70.85894	67.79722	75.53606	84.76018	78.9236	79.56622	0.895269
unit_04	50.4218	46.92915	68.86849	40.93937	35.13514	46.75757	50.35727	30.64141	11.35908	66.03053	90.65229	88.09653	29.42132	15.1433	53.39819	32.23693	44.02728	34.65555	5.174625
unit_05	44.19753	49.03165	49.52551	54.97095	28.37838	77.25748	79.95713	42.40332	4.67175	61.57761	57.62663	41.15286	38.85673	51.64526	49.86237	54.12834	48.58096	46.25948	-1.16444
unit_06	83.76485	91.93819	84.69283	95.46467	85.13514	86.58878	88.46016	55.83927	41.8366	80.40712	83.91566	88.82568	76.8501	39.20628	95.83055	67.49376	75.90923	65.7643	21.91873
unit_07	81.96521	94.98874	87.76721	83.36774	83.78378	85.96478	85.67345	78.08227	68.93955	85.24173	52.04831	89.48529	97.59145	77.41804	81.47306	67.55246	77.20703	64.42384	51.72729
unit_08	42.62445	39.68875	48.31945	42.05639	14.86486	56.62558	85.7985	59.06759	3.74025	62.84987	49.03897	80.97789	62.97883	57.35152	58.06123	73.55683	64.05617	67.18223	-7.12466
unit_09	14.07142	20.07524	21.75318	27.56527	4.054054	35.14103	30.01072	23.71567	2.338286	5.852417	0	51.04902	20.01337	28.43122	29.15764	24.02746	1.54371	23.47869	-9.97445
unit_10	75.18949	84.03577	74.31619	78.62692	82.43243	85.64292	77.42051	74.95738	56.81448	82.31552	97.70157	88.72667	71.48233	47.88481	87.19858	62.32376	73.72726	67.50337	55.74589
unit_11	32.26696	43.2165	44.70589	37.19947	36.48649	50.62543	37.44194	20.6722	14.37654	14.50382	18.77309	62.11208	8.115095	26.81464	16.14396	16.65127	25.78795	23.91937	4.180333
unit_12	28.65253	36.99057	37.58512	28.27211	16.21622	67.09783	20.29296	43.85691	3.587114	17.43003	27.13678	64.60119	25.28294	33.92418	40.70152	45.5555	45.42924	29.06799	-1.08793
unit_13	52.87184	51.70895	73.20888	63.44682	37.83784	63.07038	72.57949	43.77312	16.57781	65.52163	71.26316	92.485	44.98063	18.61908	30.85267	43.56508	48.57635	46.60544	2.340609
unit_14	51.04856	63.09581	44.19637	58.26076	29.72973	71.44983	70.09646	67.06145	1.897638	82.1883	87.68207	92.00453	67.32887	67.43946	63.53982	70.92873	70.43806	62.48435	-0.76633
unit_15	34.85376	47.72978	56.62508	44.40937	32.43243	57.6773	59.32476	27.60826	4.053746	52.92621	40.48993	49.25926	27.02927	10.86446	20.45752	23.56995	36.85114	16.99086	0.549058
unit_16	37.0723	24.44941	58.35777	26.02802	20.27027	60.88808	35.209	16.81652	3.121807	34.86005	64.84354	80.25007	6.799115	10.49406	24.06592	16.92199	24.39666	12.03045	1.164991
unit_17	39.56225	46.02418	56.12137	45.27782	28.37838	68.26128	52.51876	48.85191	27.42938	58.14249	84.55652	77.43712	26.53616	19.31561	53.56084	33.3278	53.42522	38.95555	4.005766
unit_18	74.28409	67.76948	91.12248	65.64764	66.21622	56.41531	73.99071	34.76793	16.07615	73.66412	74.27711	75.38808	54.38461	56.74737	45.2068	45.60114	46.25062	23.1991	16.31308
unit_19	60.05553	67.31844	89.77119	39.92578	72.97297	56.51303	63.36191	39.37501	11.93701	30.66158	32.03892	39.65284	61.89551	62.36656	55.06888	59.40538	40.89081	35.82888	10.95296
unit_20	56.84691	64.66887	87.60658	59.48843	63.51351	57.36542	62.1472	22.73874	11.45861	40.96692	64.86818	77.30235	31.74674	35.4416	57.10097	37.26344	36.9715	30.83934	9.549769
unit_21	91.48768	94.72543	89.3935	94.39186	90.54054	78.67368	86.90604	69.87254	34.43415	79.51654	78.62573	84.84185	75.33013	67.51885	92.40134	78.60982	76.73897	68.25103	50.85717
unit_22	96.38587	97.68401	91.55197	94.32477	95.94595	80.47952	80.52876	91.86313	73.65405	91.22137	69.87102	88.6376	100	100	96.7612	95.38438	96.14877	89.43152	43.31701
unit_23	67.17821	76.29025	73.31525	80.85748	71.62162	78.70184	73.59771	39.09833	7.625629	65.01272	85.70059	76.89969	33.80345	39.22455	68.37796	62.41142	66.95126	38.41913	0.698615
unit_24	56.76761	46.52411	59.50324	44.46297	33.78378	78.50957	76.86674	65.86412	46.34481	49.61832	53.03064	80.83325	70.11912	46.17288	57.11401	68.62243	54.02696	59.71479	39.65368
unit_25	28.56637	37.92775	40.99959	39.20415	28.37838	71.61415	48.92819	19.16471	1.714363	28.49873	43.40291	59.12562	11.69958	48.06362	43.09101	30.66564	56.03964	28.47484	-11.0082

Columns: indicators (normalised values 0-100)

Multiple imputation (MI)

Implies creating ***m-complete data sets*** by imputing *m*-values for each missing cell

Observed values in the new data sets are the same, but missing values are filled in with a distribution of imputations ***reflecting the uncertainty about the missing data***

The *m*-estimates can be combined (e.g. averaging them)

Allows statistical inference on variances, standard errors and confidence intervals of the point estimates

e.g. Amelia II software (https://gking.harvard.edu/amelia)

Combines EM algorithm with a bootstrap approach

Fills in missing values without changing relationships in the data

Enables the inclusion of all the observed data in partially missing rows

Special features for time-series-cross-section data

Works from the ***R*** command line or via a graphical user interface

Pros: often unbiased with data NMAR

Cons: difficult to communicate, computational-intensive (but increasingly automatized in statistical software)

Constructing Composite Indicators: Rules of thumb and takeaways

- ***Imputations often unreliable if*** data set contains ***more than 40% of missing values***
- @***indicator-level***: at least 50% of units should have valid data for that indicator – otherwise drop indicator and consider an alternative one
- @***unit-level***: at least 65-75% of the indicators for the unit should have valid data (threshold to be applied at pillar (dimension) level and not only at framework level!) – otherwise exclude unit from pillar/index calculations (*Step 6 - Aggregation*)
- Consider ***pros and cons and implications*** of different imputation methods, and avoid as much as possible using different methods for different indicators (although at times it might be unavoidable!)

Constructing Composite Indicators: Rules of thumb and takeaways

- EM algorithm estimates are based on available data/information; but ***if variables are not intercorrelated*** and correlations not good enough (e.g. 0.6 aprox), ***you can't make a good prediction using EM!***
- ***kNN*** is not based on correlations, only on distances between recipients and donors; therefore, it ***is handy when correlations are poor***
- When using kNN, search for really "close" donors (i.e. ***keep number of neighbours low*** (e.g. k = 3) to account for "local effects")
- kNN does not work well when number of inputs is large, due to "dimensionality" problems (i.e. ***run kNN*** computing distances ***within pillar/dimension***); ***EM (and MI) should also be run separately by pillars/dimensions***, and not for the whole dataset at once

Constructing Composite Indicators: Rules of thumb and takeaways

- ***Imputation is usually performed after normalisation:*** min-max normalisation helps to easily identify out-of-bound values when using EM, and having all data in a common meaningful scale helps to give indicators the same influence when computing distances and identifying neighbours using kNN
- ***Ignoring missing values*** (i.e. shadow imputation) ***is in itself a simple treatment***: it fills in the missing value with the mean of the values observed for that unit on the other variables in the pillar!
- ***Shadow imputation*** (by rows) ***and mean imputation*** (by columns) provide ***clear incentives not to report low performance***
- ***Assess*** the sensitivity of rankings to different imputation methods (*Step 8 – **Robustness & sensitivity***)

Missing values – suggested readings

- Beretta, L. and A. Santaniello (2016). Nearest neighbor imputation algorithms: a critical evaluation. BMC Medical Informatics and Decision Making, 16 (Suppl 3):74.
- Chen, Y. & M.R. Gupta, 2010 EM Demystified: An Expectation-Maximization Tutorial. Department of Electrical Engineering. University of Washington
- Dondersa et al., 2006. Review: A gentle introduction to imputation of missing values. Journal of Clinical Epidemiology. 59: 1087-1091
- Enders, C. K., 2010, Applied Missing Data Analysis. The Guilford Press. Inc: New York, London
- He, Y., 2010, Missing Data Analysis Using Multiple Imputation: Getting to the Heart of the Matter. Circ Cardiovasc Qual Outcomes. 3(1): 98
- Graham, J. W., 2012. Missing data: Analysis and design. New York: Springer.
- Kang, H., 2013, The prevention and handling of the missing data. The Korean Journal of Anesthesiology. 64(5): 402–406.
- Little, R.A & Rubin, D.A., 2002. Statistical Analysis with missing data. Second edition. Hoboken, John Wiley & Sons.
- Schafer, J. L. & Graham, J.W., 2002, Missing Data: Our View of the State of the Art Psychological Methods. 7(2):147–177

Missing values – suggested software

Software Package	link
Freeware	
Amelia	http://gking.harvard.edu/amelia
CAT	http://cat.texifter.com/ (for categorical data)
EMCOV	https://methodology.psu.edu/publications/books/missing
NORM	https://methodology.psu.edu/publications/books/missing
MICE	http://www.stefvanbuuren.nl/mi/index.html
PAN	http://stat.ethz.ch/~maechler/adv_topics_compstat/MissingData_Imputation.html (Free with R, commercial with S-Plus, for clustered data, including longitudinal data).
Commercial Software	
AMOS	https://www.ibm.com/us-en/marketplace/structural-equation-modeling-sem
EQS	http://www.mvsoft.com
HLM	http://www.ssicentral.com/hlm/index.html
LISREL	http://www.ssicentral.com/index.html
Mplus	http://www.statmodel.com
SAS	https://www.sas.com/it_it/home.html
SOLAS	https://www.statcon.de/shop/en/software/statistics/solas
S-Plus	http://www.solutionmetrics.com.au/products/splus/default.html
SPSS	http://www-01.ibm.com/software/analytics/spss/products/statistics/modules/
Stata	http://www.stata.com, *mi* command; installing *ice* or *mvis*
	Source: updated from Acock, 2005

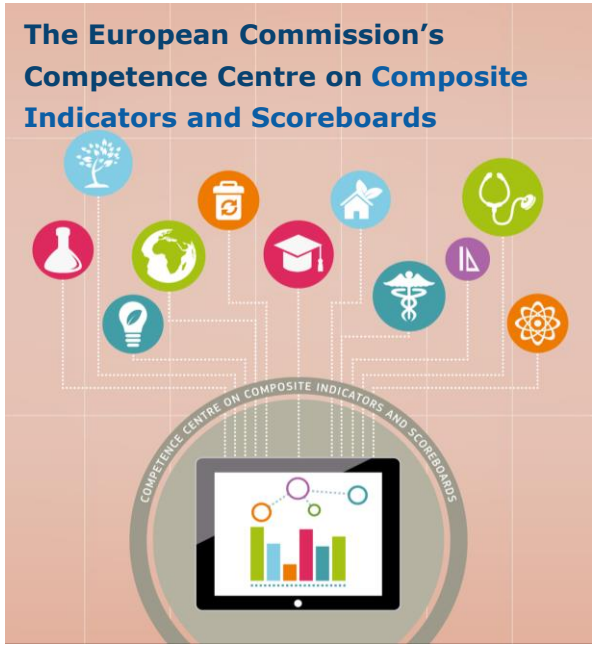
THANK YOU

Any questions?

Welcome to email us at: jrc-coin@ec.europa.eu

COIN in the EU Science Hub
https://ec.europa.eu/jrc/en/coin

COIN tools are available at:
https://composite-indicators.jrc.ec.europa.eu/



European Commission