The European Commission's science and knowledge service

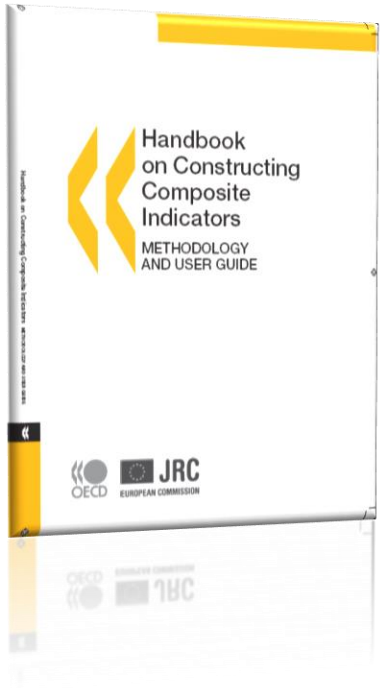
Joint Research Centre

European Commission

Step 3: Missing Data

Maria Del Sorbo

COIN 2017 - 15th JRC Annual Training on Composite Indicators & Scoreboards
06-08/11/2017, Ispra (IT)



Step 10. Presentation & dissemination

Step 9. Association with other variables

Step 8. Back to the indicators

Step 7. Robustness & sensitivity

Step 6. Weighting & aggregation

Step 5. Normalization of data

Step 4. Multivariate analysis

Step 3. Data treatment (missing, outliers**)**

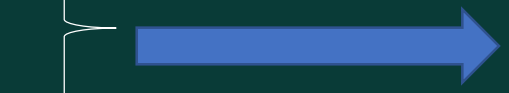
Step 2. Selection of indicators

Step 1. Developing the framework

Why is it important to treat missing values?

Ignoring missing data may:

- reduce the statistical power of the data
- generate biased estimates
- reduce the representativeness of the samples



baseless CI

The best possible method of handling the missing data is to prevent the problem by well-planning the study and collecting the data carefully (Kang, 2013).

What is our aim today?

To learn about:

- Ways to define and understand missing data
- How to distinguish between types of missing data
- Techniques for handling missing data, (dis) advantages
- Operationalizing with some examples

What are missing data?

‘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)

Missing data analysis stages

1. *What type of data am I dealing with?* (variable, e.g., categorical versus continuous; source, e.g., official statistics, surveys, etc.)

2. Identify patterns/rationale for missing and recode correctly

3. *What is distribution of the missing data?*

4. Choose the most suitable method of data analysis

STAGE 1: Discuss ways to define and understand missing data

- Attrition due to social/natural processes, e.g., *school graduation, dropout, death*
- Skip pattern in survey, e.g., *some questions are directed only for respondents (firms) who are innovative (CIS examples)*
- Random data collection issues
- Respondent refusal/Non-response, i.e. the required information is not obtained from the persons selected in the sample

STAGE 1: Skip pattern in survey

E.g. Community Innovation Survey 2014 – firms data

2.1 During the three years 2012 to 2014, did your enterprise introduce:

	Yes *1*	No *0*	
Goods innovations: New or significantly improved goods (exclude the simple resale of new goods and changes of a solely aesthetic nature)	☐	☐	*INPDGD*
Service innovations: New or significantly improved services	☐	☐	*INPDSV*

If no to all options, go to section 3

Otherwise, go to question 2.2

2.2 Who developed these product innovations?

Tick all that apply

	Goods innovations		Service innovations	
Your enterprise by itself	☐	*INITGD*	☐	*INITSV*
Your enterprise together with other enterprises or organisations*	☐	*INTOGD*	☐	*INTOSV*
Your enterprise by adapting or modifying goods or services originally developed by other enterprises or organisations*	☐	*INADGD*	☐	*INADSV*
Other enterprises or organisations	☐	*INOTHGD*	☐	*INOTHSV*

**: Include independent enterprises plus other parts of your enterprise group (subsidiaries, sister enterprises, head office, etc.). Organisations include universities, research institutes, non-profits, etc.*

2: Distinguish types of missing data

Three mechanisms of missingness:

- Missing completely at random, MCAR =>completely unsystematic missingness on the variable
- Missing at random, MAR => missingness pattern related to other variables
- Missing Not at random, MNAR=> missingness pattern related to the values of the variable itself

Which missingness mechanism does this example refer to MCAR, MAR, MNAR?

	Complete data		Incomplete data	
Cases	Age	Creative thinking (score 0-100)	Age	Creative thinking (score 0-100)
1	30	93	30	93
2	31	81	31	.
3	32	91	32	91
4	33	100	33	100
5	34	95	34	95
6	35	99	35	.
7	36	98	36	98
8	37	50	37	50
9	38	47	38	47
10	39	94	39	94
11	40	80	40	.

Well-done, it is MCAR!

Which missingness mechanism does this example refer to MCAR, MAR, MNAR?

Cases	Age	Creative thinking (score 0-100)	Age	Creative thinking (score 0-100)
1	30	93	30	.
2	31	81	31	.
3	32	91	32	.
4	33	100	33	.
5	34	95	34	.
6	35	99	35	.
7	36	98	36	98
8	37	50	37	50
9	38	47	38	47
10	39	94	39	94
11	40	80	40	93

Well-done, it is MAR!

Which missingness mechanism does this example refer to MCAR, MAR, MNAR?

Cases	Age	Creative thinking (0-100)	Age	Creative thinking (0-100)
1	30	93	30	93
[…]	[…]	[…]	[…]	[…]
6	35	99	35	99
7	36	98	36	98
8	37	50	37	.
9	38	47	38	.
10	39	94	39	94
11	40	80	40	80
12	41	35	41	.
13	42	97	42	97
14	43	70	43	70
15	44	49	44	.
16	45	65	45	65

Well-done, it is MNAR!

Identified the missing data distribution, What next?

STAGE 3: Techniques for handling missing data

Most suitable technique…

- To prevent missing values
- Assumption MCAR, or MAR; Little´s MCAR test
- Use all the available information about why data are missing
- It depends on the missing data rate (e.g. very small rate about 2%=> mean imputation)

STAGE 3: Techniques for handling missing data

Most suitable technique…

Select the imputation method to generate the least biased estimates (it depends upon the research field, e.g., econometricians, statisticians):

- Deletion: listwise, pairwise
- Ignore the missing data
- Single imputation: mean/median/mode substitution, hot deck, single regression, etc.
- Model-based (Expectation Maximization Maximum Likelihood, Multiple Imputation)

STAGE 3: Deletion methods

STAGE 3: deletion methods, listwise

Economy	Exp_Edu	Govt_Exp_Edu	SchooLifExpectancy	Tertiaryenrolment	SciEngGraduates
Bolivia	7.3	18.4	.	.	.
Canada	5.3	18.3	.	.	.
Croatia	4.6	.	**15.3**	**69.5**	**23.8**
Germany	4.9	23.7	17.3	68.3	.
Greece	.	.	17.8	113.9	28.7
Iceland	7.8	**18.3**	**19.6**	**81.3**	**15.6**
Ireland	5.3	26	19	77.6	23.8
Norway	7.4	25.8	17.7	76.7	20
Average	**6.09**	**21.75**	**17.78**	**81.22**	**22.38**

Source: author elaboration, 2016 GII data sub-sample

Advantage: the remaining dataset is complete
Disadvantage: reduced sample size and power, caused by the loss of the incomplete cases.

STAGE 3: deletion methods, pairwise

Economy	Exp_Edu	Govt_Exp_Edu	SchooLifExpectancy	Tertiaryenrolment	SciEngGraduates
Bolivia	7.3	18.4	.	.	.
Canada	5.3	18.3	.	.	.
Croatia	4.6	.	**15.3**	**69.5**	**23.8**
Germany	4.9	23.7	17.3	68.3	.
Greece	.	.	17.8	113.9	28.7
Iceland	7.8	**18.3**	**19.6**	**81.3**	**15.6**
Ireland	5.3	26	19	77.6	23.8
Norway	7.4	25.8	17.7	76.7	20

Source: author elaboration, 2016 GII data sub-sample

Advantage: more data conservative compared to listwise
Disadvantage: the sample size will remain the same for some analyses and will be reduced for others; inconsistency of the sample size can lead to problems in computing standard errors

What is ‘imputation’?

‘Imputation of missing data on a variable is replacing that missing by a value that is drawn from an estimate of the distribution of this variable’ (Dondersa et al., 2006)

STAGE 3: Single imputations

STAGE 3: Single imputations, mean imputation

Economy	Exp_Edu	Govt_Exp_Edu	SchooLifExpectancy	Tertiaryenrolment	SciEngGraduates
Bolivia	7.3	18.4	**17.78**	**81.22**	**22.38**
Canada	5.3	18.3	**17.78**	**81.22**	**22.38**
Croatia	4.6	**21.75**	15.3	69.5	23.8
Germany	4.9	23.7	17.3	68.3	**22.38**
Greece	**6.09**	**21.75**	17.8	113.9	28.7
Iceland	7.8	18.3	19.6	81.3	15.6
Ireland	5.3	26	19	77.6	23.8
Norway	7.4	25.8	17.7	76.7	20
Average	**6.09**	**21.75**	**17.78**	**81.22**	**22.38**

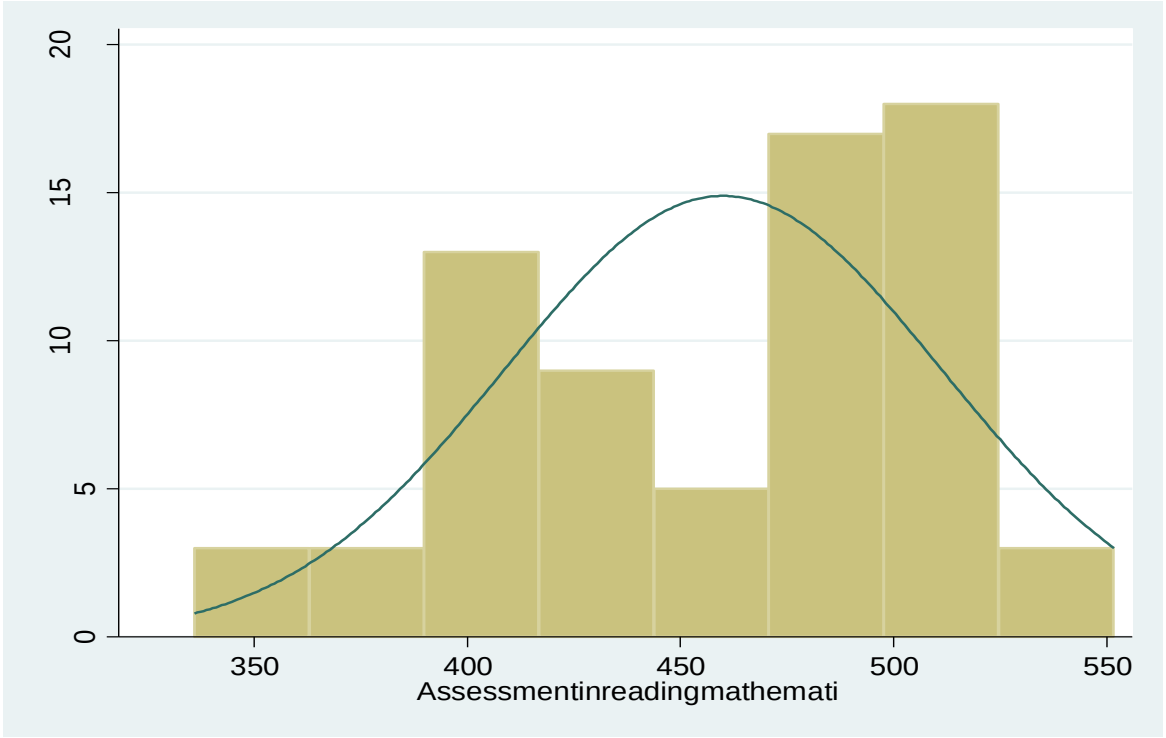
Source: author elaboration, 2016 GII data sub-sample

Mean imputation: replaces the missing values with the arithmetic average of the same variable

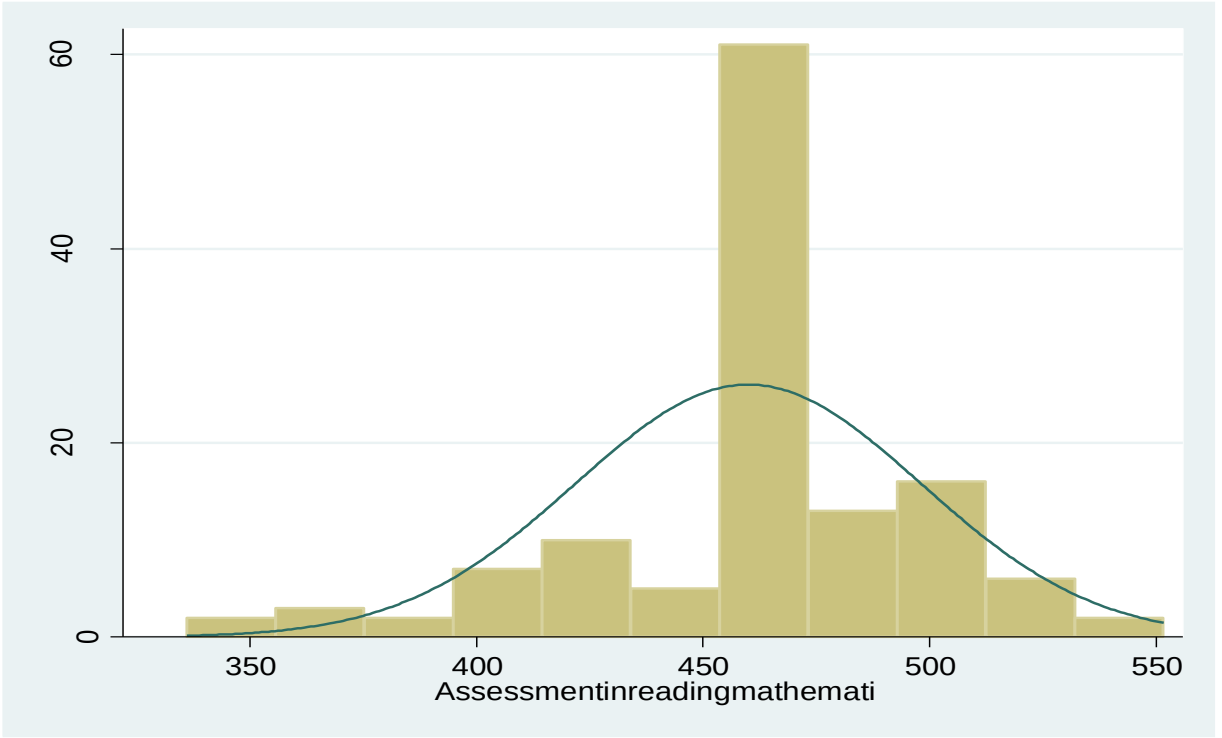
3: Single imputation, mean

Data frequency distribution

With missing values



After mean imputation



Source: author elaboration, 2016 GII data

Advantage: simple and straightforward
Disadvantage: distorts distribution, attenuates variance=>biased estimates, it modifies relationships between variables

European Commission

STAGE 3: single imputations, hot deck

Hot deck imputation is a non-parametric regression by means of the k-Nearest Neighbor estimator. It replaces missing values with observed values having some *mathematical* similarities

Advantages :imputed values are real values, it preserves the original structure of the data, NN being non parametric does not require explicit models to relate y and x

Disadvantages: computational costs may be high, instance-based learning algorithm

STAGE 3: single imputations, hot deck

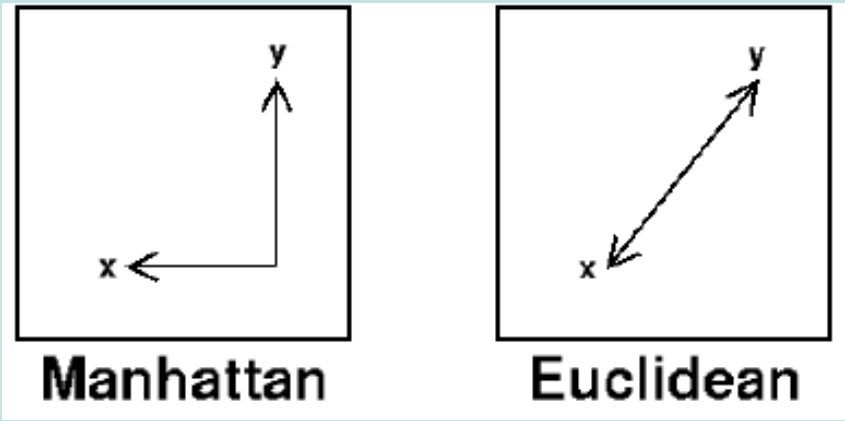
Minkowski distances

- **Manhattan distance**: the distance between two points in a grid based on a strictly horizontal and/or vertical path
- **Euclidean distance** is the "ordinary“ straight line distance between two points in Euclidean space
- **Supremum distance** is the maximum absolute value of difference between two points in an infinite space, also called ∞ metric

STAGE 3: single imputations, hot deck kNN

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

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

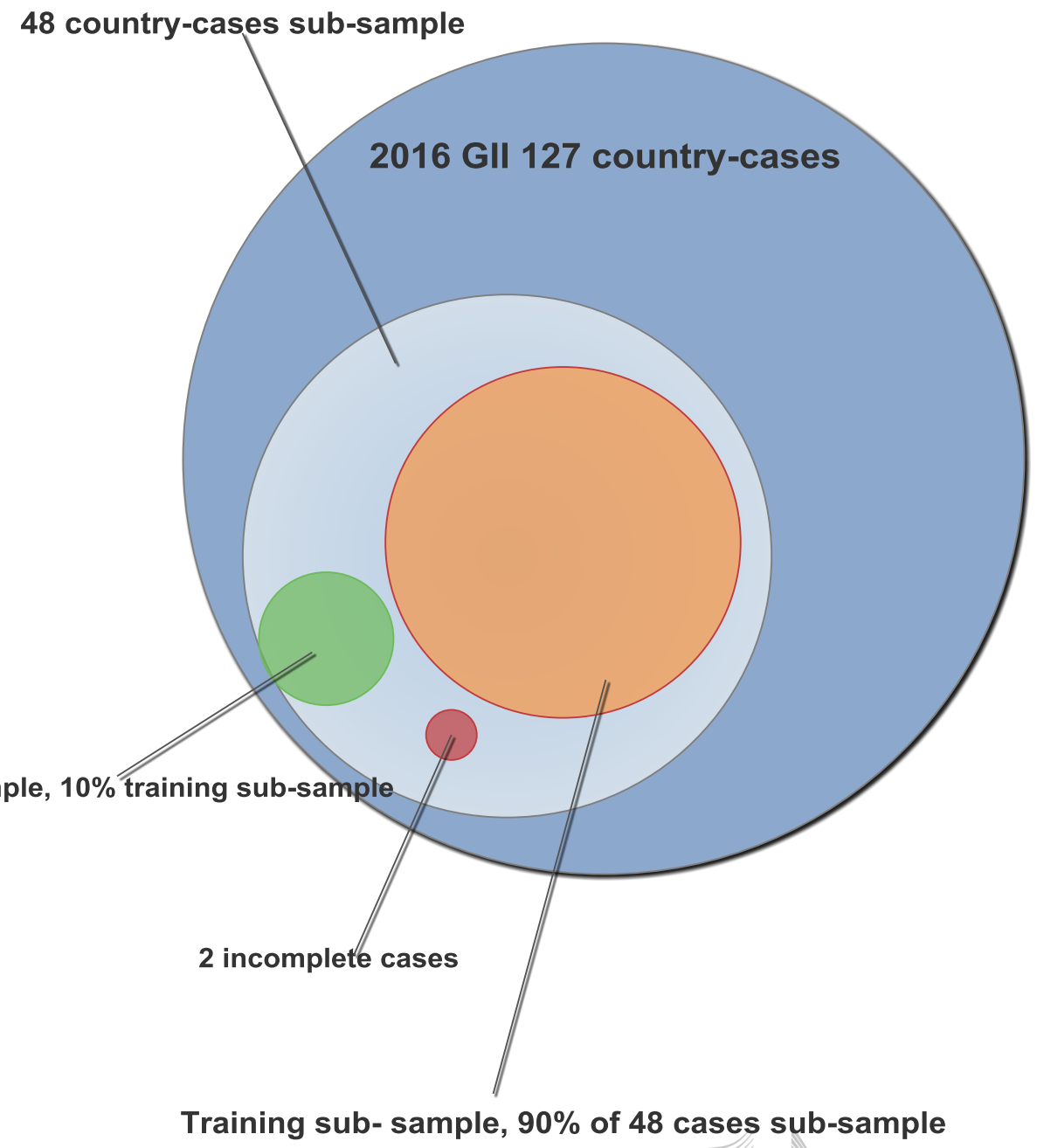


- **Supremum distance** $d=((x_1 , y_1), (x_2 , y_2)) = max(|x_1 - x_2|, |y_1 - y_2|)$

STAGE 3: hot deck imputation, how to do it?

Case study: 2016 Global Innovation sub-sample

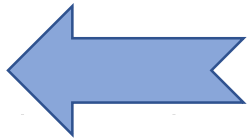
- Missing variable to impute: School Life Expectancy (predicted variable)
- Observed variables used: Expenditure in Education and Government Expenditure (predictors)
- 2 country-cases to impute
- Assumption: our data are MCAR, or MAR
- Min max standardization



STAGE 3: hot deck techniques comparisons

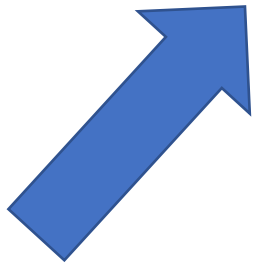
2016 GII sub-sample		PredictorX	PredictorY	Predicted	Coding			
Type of sample	Economy	Expenditureoneducation	Governmentexpenditureoneducat	Schoollifeexpectancy	NoMissing	Manhattan	Euclidean	Supremum
Training sub-sample	Austria	0.55	0.62	0.56	1	0.244	0.174	0.138
Training sub-sample	Belgium	0.67	0.91	1.00	1	0.657	0.486	0.429
Training sub-sample	Brazil	0.61	0.45	0.48	1	0.198	0.170	0.167
Training sub-sample	Bulgaria	0.32	0.47	0.43	1	0.136	0.122	0.121
Training sub-sample	Chile	0.42	0.27	0.61	1	0.231	0.217	0.216
Training sub-sample	[…]	[…]	[…]	[…]	[…]	0.254	0.202	0.193
Training sub-sample	Lebanon	0.09	0.00	0.00	1	0.830	0.594	0.481
Training sub-sample	Lithuania	0.39	0.34	0.61	1	0.187	0.148	0.141
Training sub-sample	Luxembourg	0.32	0.39	0.32	1	0.216	0.154	0.121
Training sub-sample	Malaysia	0.45	0.36	0.20	1	0.139	0.125	0.124
Training sub-sample	Malta	0.95	1.00	0.44	1	1.031	0.729	0.516
Training sub-sample	Mexico	0.48	0.31	0.25	1	0.221	0.182	0.176
	Netherlands	0.55	0.53	0.80	1	0.155	0.117	0.106
	New Zealand	0.65	0.46	0.92	1	0.232	0.213	0.212
	Spain	0.35	0.48	0.77	1	0.097	0.091	0.091
	Sweden	0.86	0.54	0.81	1	0.479	0.428	0.424
Testing	Switzerland	0.47	0.58	0.56	1			
	Tunisia	0.65	0.53	0.42	1			
	Turkey	0.42	0.26	0.60	1			
	United Kingdom	0.58	0.50	0.77	1			
	United States of America	0.44	0.48	0.61	1			
To be imputed	Bolivia, P.S.	0.80	0.36	.	0			
To be imputed	Canada	0.50	0.36	.	0			

X5 country-cases



STAGE 3: Minkowski techniques Cross-validation

	Manhattan					Euclidean					Supremum				
	K1	K2	K3	K4	K5	K1	K2	K3	K4	K5	K1	K2	K3	K4	K5
Switzerland	0.705	0.636	0.621	0.605	0.643	0.705	0.636	0.621	0.605	0.582	0.705	0.636	0.621	0.588	0.582
Tunisia	0.795	0.801	0.769	0.807	0.759	0.795	0.801	0.769	0.719	0.7	0.795	0.801	0.723	0.719	0.695
Turkey	0.614	0.494	0.492	0.432	0.425	0.614	0.494	0.492	0.469	0.425	0.614	0.494	0.492	0.469	0.425
UK	0.602	0.614	0.674	0.736	0.684	0.602	0.761	0.716	0.736	0.709	0.602	0.602	0.708	0.688	0.709
USA	0.602	0.602	0.765	0.651	0.607	0.602	0.602	0.659	0.651	0.641	0.602	0.602	0.659	0.651	0.641
MAPE	0.283	0.29	0.3	0.279	0.276	0.283	0.252	0.255	0.225	0.226	0.283	0.293	0.235	0.232	0.224



STAGE 3: Minkowski techniques Cross-validation

Mean Absolute Percentage Error, MAPE

$$\text{MAPE}\sum_i^n = |(Oi - P_i)/Oi|/\text{n}$$

	Manhattan	Euclidean	Supremum
Switzerland	0.153	0.131	0.125
Tunisia	0.870	0.800	0.776
Turkey	0.192	0.179	0.179
UK	0.143	0.088	0.143
USA	0.071	0.043	0.043
Average of MAPEs (Kn)	**0.286**	**0.248**	**0.253**

STAGE 3: Minkowski techniques Cross-validation

Mean Absolute Percentage Error, MAPE=> K4

	Manhattan					Euclidean					Supremum				
	K1	K2	K3	K4	K5	K1	K2	K3	K4	K5	K1	K2	K3	K4	K5
Switzerland	0.705	0.636	0.621	0.605	0.643	0.705	0.636	0.621	0.605	0.582	0.705	0.636	0.621	0.588	0.582
Tunisia	0.795	0.801	0.769	0.807	0.759	0.795	0.801	0.769	0.719	0.7	0.795	0.801	0.723	0.719	0.695
Turkey	0.614	0.494	0.492	0.432	0.425	0.614	0.494	0.492	0.469	0.425	0.614	0.494	0.492	0.469	0.425
UK	0.602	0.614	0.674	0.736	0.684	0.602	0.761	0.716	0.736	0.709	0.602	0.602	0.708	0.688	0.709
USA	0.602	0.602	0.765	0.651	0.607	0.602	0.602	0.659	0.651	0.641	0.602	0.602	0.659	0.651	0.641
MAPE	0.283	0.29	0.3	0.279	0.276	0.283	0.252	0.255	0.225	0.226	0.283	0.293	0.235	0.232	0.224

STAGE 3: KNN- K4 imputation

	Economy	PredictorX Education Expenditure	PredictorY Government Expenditure	Predicted School Life Expectancy	No Missing	Euclidean	Euclidean K1	K2	K3	K4	K5
Training Dataset	Austria	0.55	0.62	0.56	1	0.085831204					
	[…]	[…]	[…]	[…]	1						
	France	0.53	0.60	0.59	1	0.063875198					
	Georgia	0.00	0.24	0.49	1	0.581569197					
	Hungary	0.33	0.30	0.49	1	0.316216083					
	Japan	0.27	0.55	0.49	1	0.199066733					
	Korea, R.	0.39	0.50	0.63	1	0.108598227					
	Latvia	0.44	0.56	0.57	1	0.034889176					
	The Netherlands	0.55	0.53	0.80	1	0.090218419					
	Slovakia	0.32	0.40	0.44	1	0.238692756					
	Slovenia	0.53	0.58	0.70	1	0.06087951					
Testing	**Switzerland**	**0.47**	**0.58**	**0.56**	1		0.705	0.636	0.621	0.605	0.582
	Tunisia	0.65	0.53	0.42	1						
	Turkey	0.42	0.26	0.60	1						
	United Kingdom	0.58	0.50	0.77	1						
	USA	0.44	0.48	0.61	1						

Pearson correlation Observed and KNN predicted data

Observed data - Correlations

Pearson Correlation	Expenditureoneducat	Governmentexpenditureone	Schoollifeexpecta
Expenditureoneducation	1		
Governmentexpenditureonedu	0.770	1	
Schoollifeexpectancy	**0.468**	**0.421**	1
Assessmentinreadingmathem	0.470	0.519	0.780
Pupilteacherratiosecondary	-0.123	-0.253	-0.201
Tertiaryenrolment	0.348	0.379	0.802
Graduatesinscienceandengine	0.046	-0.004	-0.205
Tertiaryinboundmobility	0.063	0.050	-0.014

Data with Predicted KNN Euclidead K4 - Correlations

Pearson Correlation	Expenditureoneducat	Governmentexpenditureone	Schoollifeexpecta
Expenditureoneducation	1		
Governmentexpenditureonedu	0.760	1	
Schoollifeexpectancy	**0.483**	**0.424**	1
Assessmentinreadingmathem	0.415	0.497	0.719
Pupilteacherratiosecondary	-0.107	-0.273	-0.178
Tertiaryenrolment	0.293	0.319	0.720
Graduatesinscienceandengine	0.123	0.027	-0.091
Tertiaryinboundmobility	0.069	0.074	0.012

Predictors significance level at 99%

3: Techniques for handling missing data: modern model-based: EM-ML

Expectation Maximization-ML: is a method to find the maximum likelihood estimator of a parameter θ of a probability distribution (Chen & Gupta, 2010).

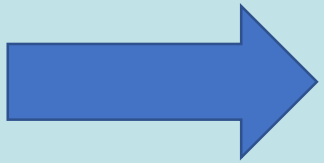
EM-ML function: $g(y \mid \int\theta) = \{x \text{ in } \Omega X\} f(x,y \mid \theta)\, dx$

Advantage: unbiased estimates, it preserves the relationships between variables
Disadvantage: it does not necessarily take into account standard errors, low missing data rate, e.g. 5%

3: Techniques for handling missing data: EM-ML

- **Assumption:** data MCAR, or MAR
- Little´s MCAR **test**

- Results may change =>Software used, number of iterations, etc.
- Correlation structure EM ML respects the real values correlation structure compared to hot deck KNN

  - *Not in our GII case study*  *counterexample*

Pearson correlation Observed and EM-ML predicted data

Observed data - Correlations

Pearson Correlation	Expenditureoneducat	Governmentexpenditureone	Schoollifeexpecta
Expenditureoneducation	1		
Governmentexpenditureonedu	0.770	1	
Schoollifeexpectancy	**0.468**	**0.421**	1
Assessmentinreadingmathem	0.470	0.519	**0.780**
Pupilteacherratiosecondary	-0.123	-0.253	-0.201
Tertiaryenrolment	0.348	0.379	0.802
Graduatesinscienceandengine	0.046	-0.004	-0.205
Tertiaryinboundmobility	0.063	0.050	-0.014

Data with EM Predicted - Correlations[a]

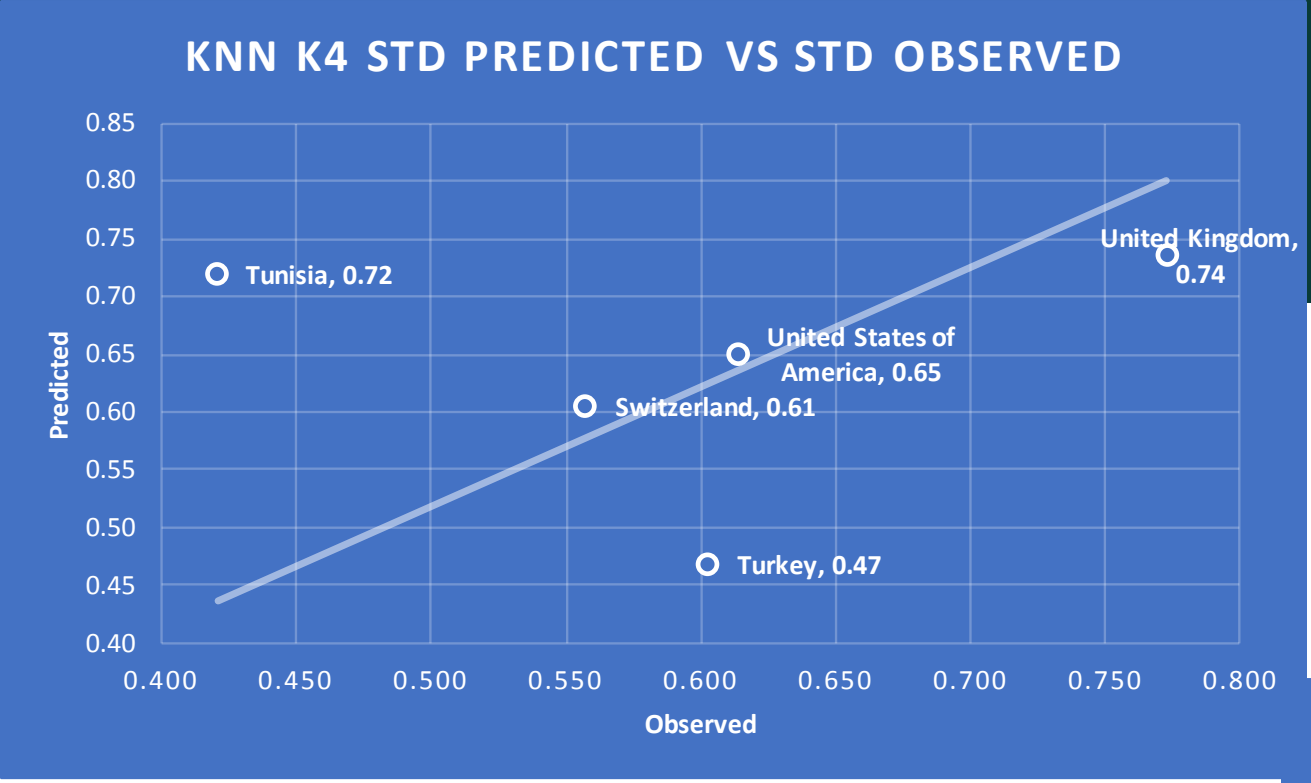
Pearson Correlation	Expenditureoneducat	Governmentexpenditureone	Schoollifeexpecta
Expenditureoneducation	1		
Governmentexpenditureonedu	0.760	1	
Schoollifeexpectancy	**0.391**	**0.367**	1
Assessmentinreadingmathem	0.415	0.497	**0.772**
Pupilteacherratiosecondary	-0.107	-0.273	-0.157
Tertiaryenrolment	0.293	0.319	0.799
Graduatesinscienceandengine	0.123	0.027	-0.344
Tertiaryinboundmobility	0.069	0.074	0.017

a. Little's MCAR test: Chi-Square = 4.482, DF = 7, Sig. = .723

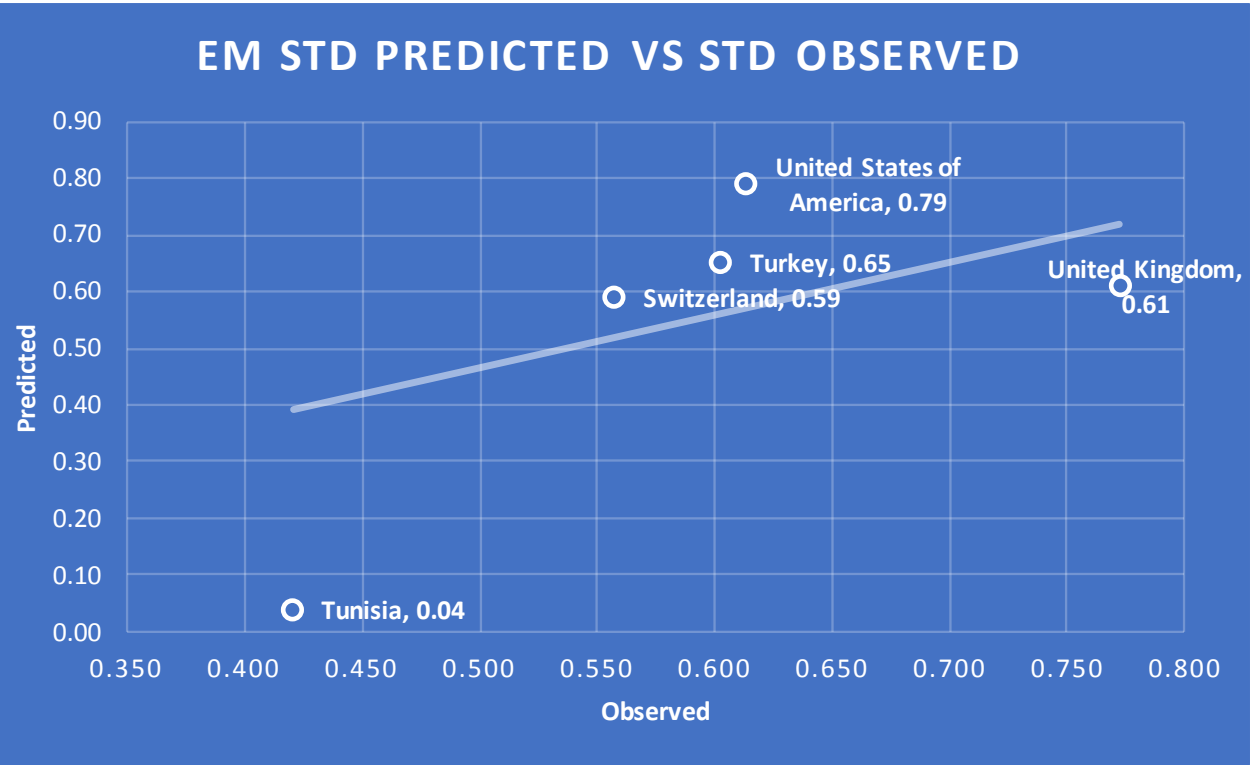
Predictors significance level at 99%

Scatter plots:

School Life Expectancy KNN K4 & EM-LM Standardized Predicted vs Standardized Observed

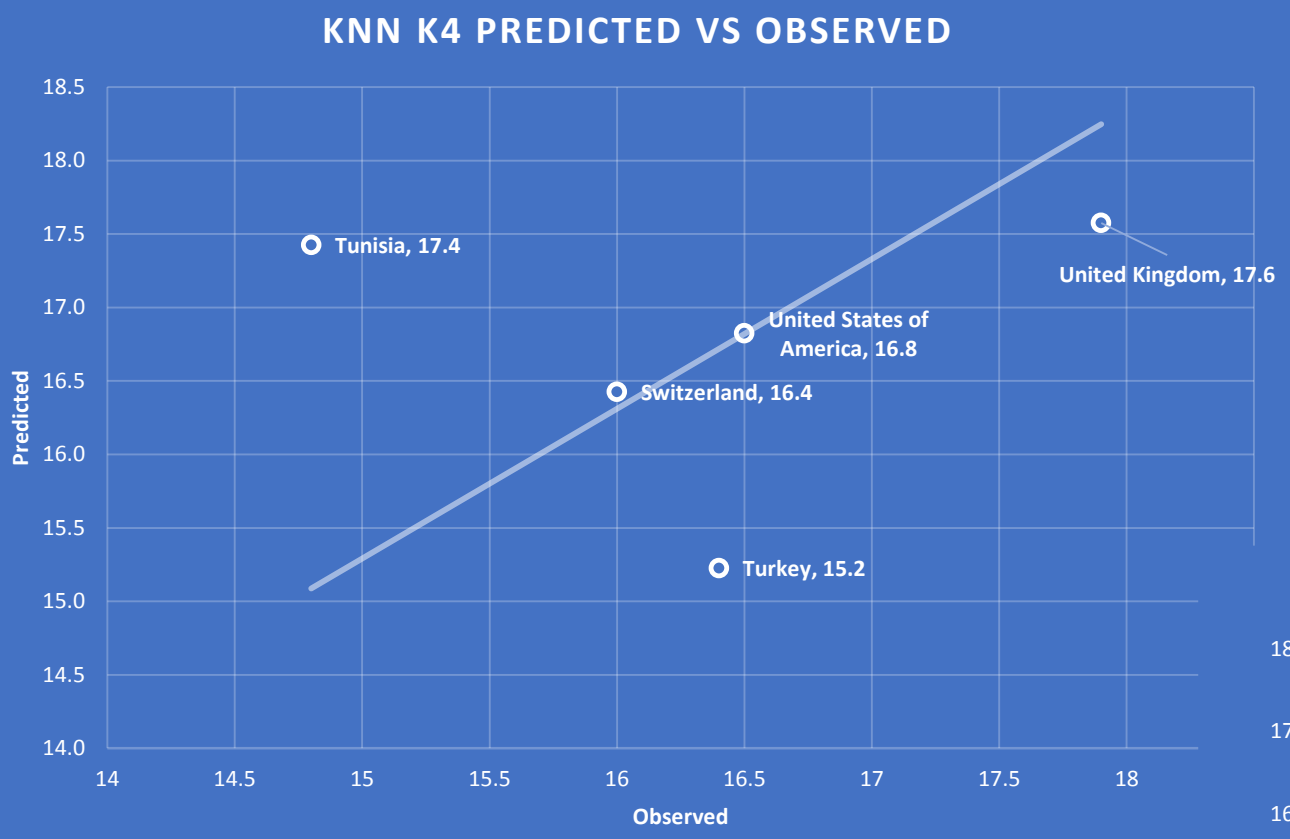


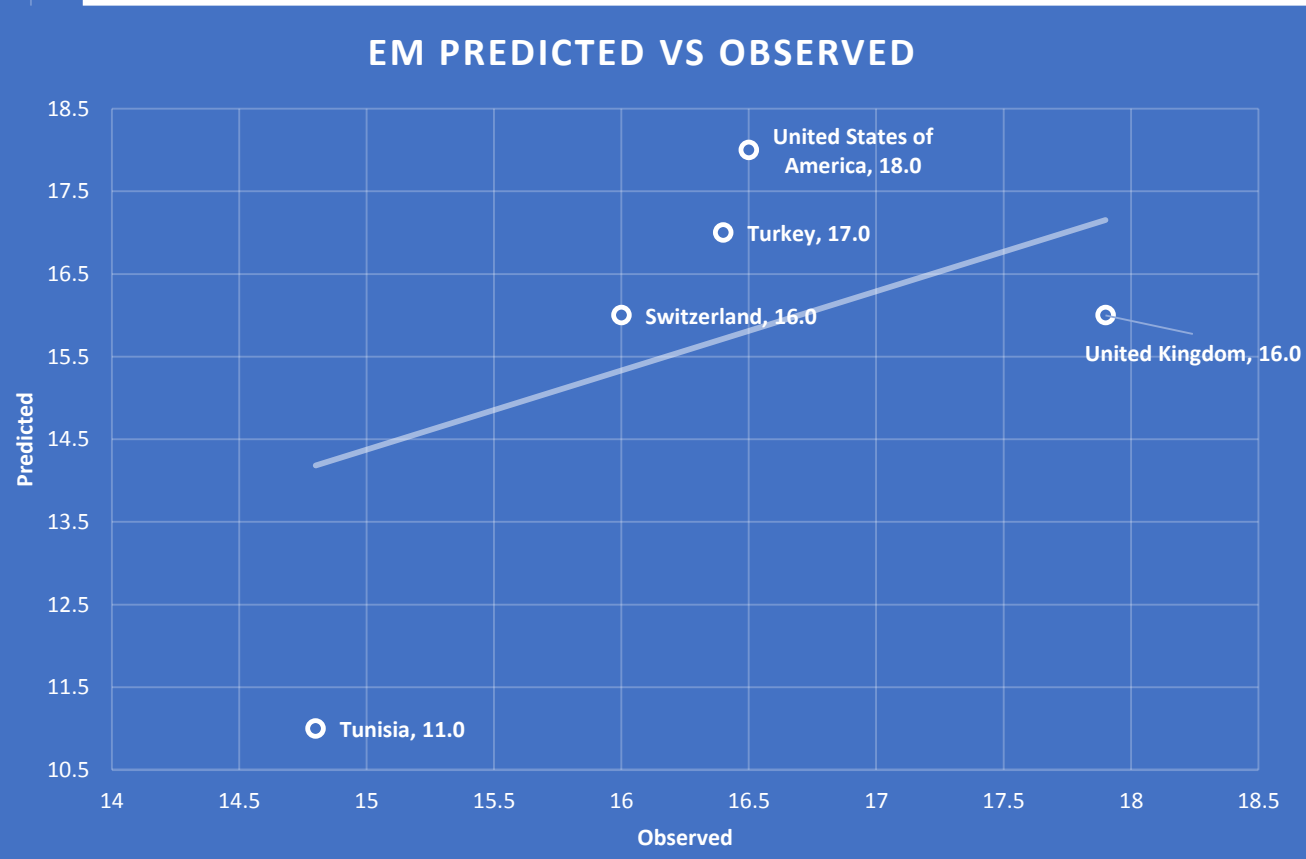
School Life Expectancy			
Economy	Observed std	KNN K4 Predicted std	EM Predicted std
Switzerland	0.56	0.61	0.59
Tunisia	0.42	0.72	0.04
Turkey	0.60	0.47	0.65
United Kingdom	0.77	0.74	0.61
United States of America	0.61	0.65	0.79



Scatter plots:

School Life Expectancy KNN K4 & EM Predicted vs Observed





School Life Expectancy

Economy	Observed std	KNN K4 Pred. std	EM Pred. std
Switzerland	16	16.4	16.0
Tunisia	14.8	17.4	11.0
Turkey	16.4	15.2	17.0
United Kingdom	17.9	17.6	16.0
United States of America	16.5	16.8	18.0

3: Disadvantages of using single imputations

Single imputation will tend to **under-estimate** the **standard errors** and thus overestimate the level of precision. Thus, single imputation gives the researcher more apparent power than the data justify…

Multiple imputation (m separate data sets are imputed) allows pooling of the parameter estimates to obtain an improved parameter estimate. (Acock, 2005)

3: Techniques for handling missing data: model-based: Multiple Imputations

MI=> different solutions for each imputation

- If the *m* **solutions** are very **similar**, this supports the imputation
- If the *m* **solutions** are significantly **different**, it is important to incorporate this uncertainty into the standard errors. Multiple imputation allows a researcher to incorporate this missing data uncertainty (Acock, 2005)
- It allows to respect the type of variable to be imputed, e.g. **categorical** vs continuous

Selected Software Packages used in working with missing values

Software Package	Selected Software Packages used in working with missing values
Freeware	**link**
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, installing ice or mvis
	Source: Acock, 2005 with author's webpage updates

References

Acock, A., C., 2005, Working with missing values. Journal of Marriage and Family. 67 (4)

Agresti, A., 2002, Categorical Data Analysis. John Wiley & Sons, Inc: Hoboken, New Jersey

Chen, Y. & M.R. Gupta, 2010 EM Demystified: An Expectation-Maximization Tutorial. Department of Electrical Engineering. University of Washington

Dempster et al., 1977, Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological).* 39(1): 1-38

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

Rubin D.B., 1976, Inferences and missing data. Biometrika. 63:581-90

Humphries, M., 2010. Missing Data & How to Deal: An overview of missing data

Little, R. J., Rubin, D., 2002. Statistical Analysis with Missing Data. John Wiley & Sons, Inc: Hoboken.

Graham, J. W., 2009. Missing Data Analysis: Making It Work in the Real World. *The Annual Review of Psychology.* 60:549-576

Graham, J. W., 2012. *Missing data: Analysis and design.* New York: Springer.

Graham, J. W., 2012. *MI automate users' guide: Steps for making NORM work with SPSS & with 2-level HLM analysis.* University Park: Penn State. Retrieved from http://methodology.psu.edu

Kang, H., 2013, The prevention and handling of the missing data. *The Korean Journal of Anesthesiology.* 64(5): 402–406.

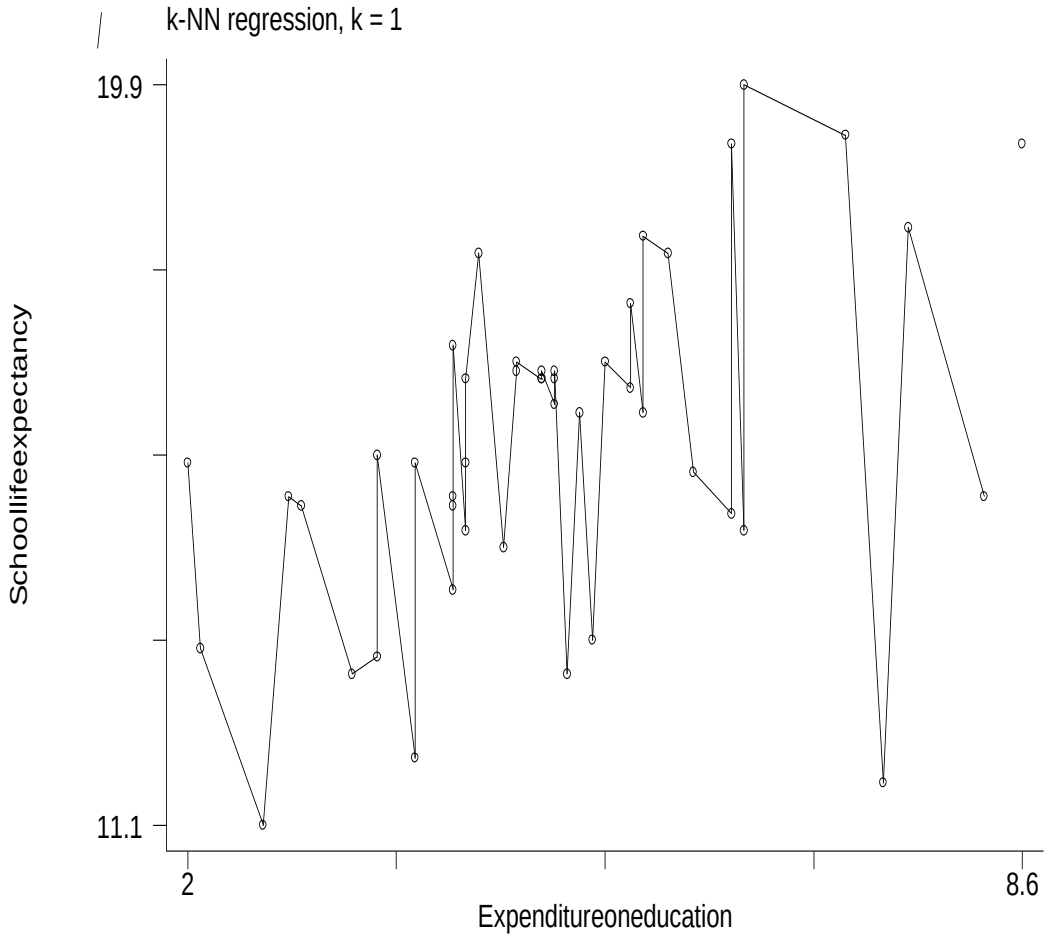
Schafer, J. L. & Graham, J.W., 2002, Missing Data: Our View of the State of the Art Psychological Methods. 7(2):147–177

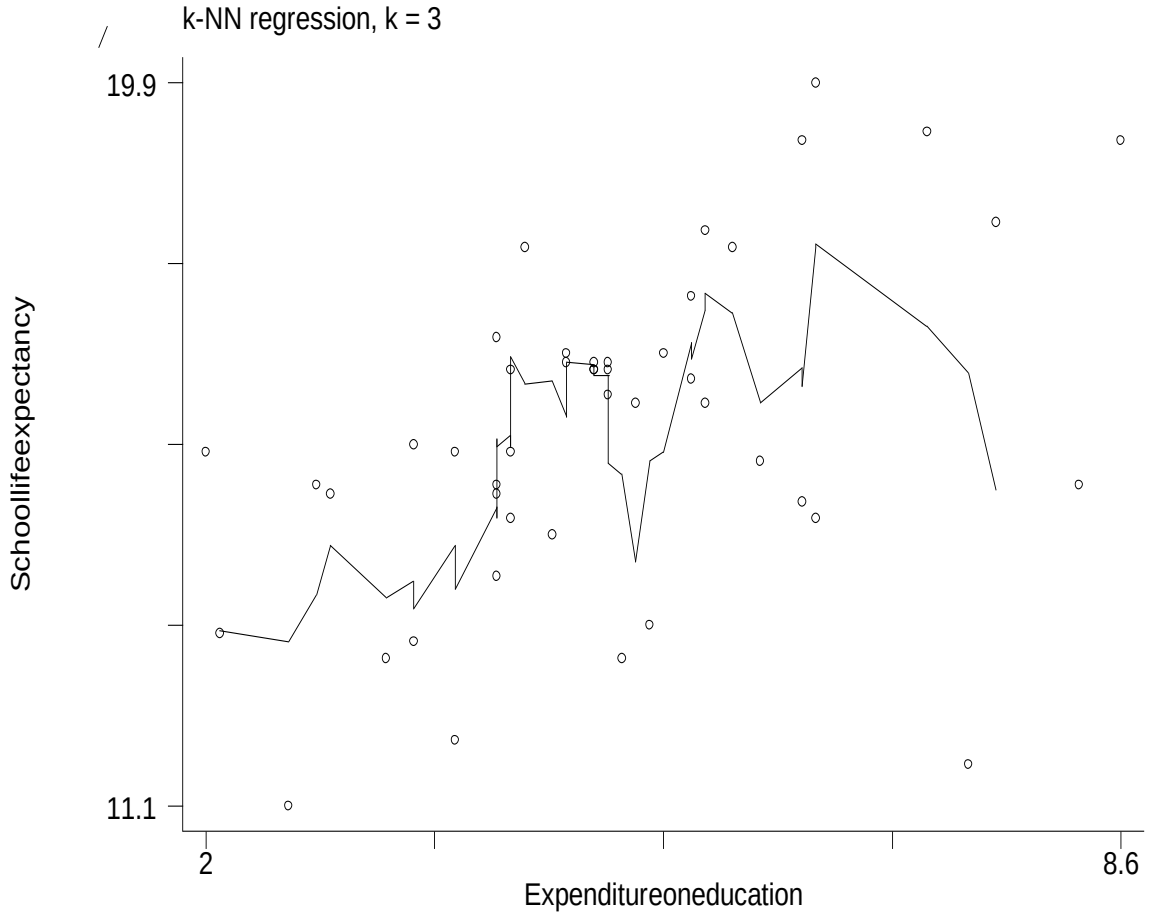
Appendix

Correlations

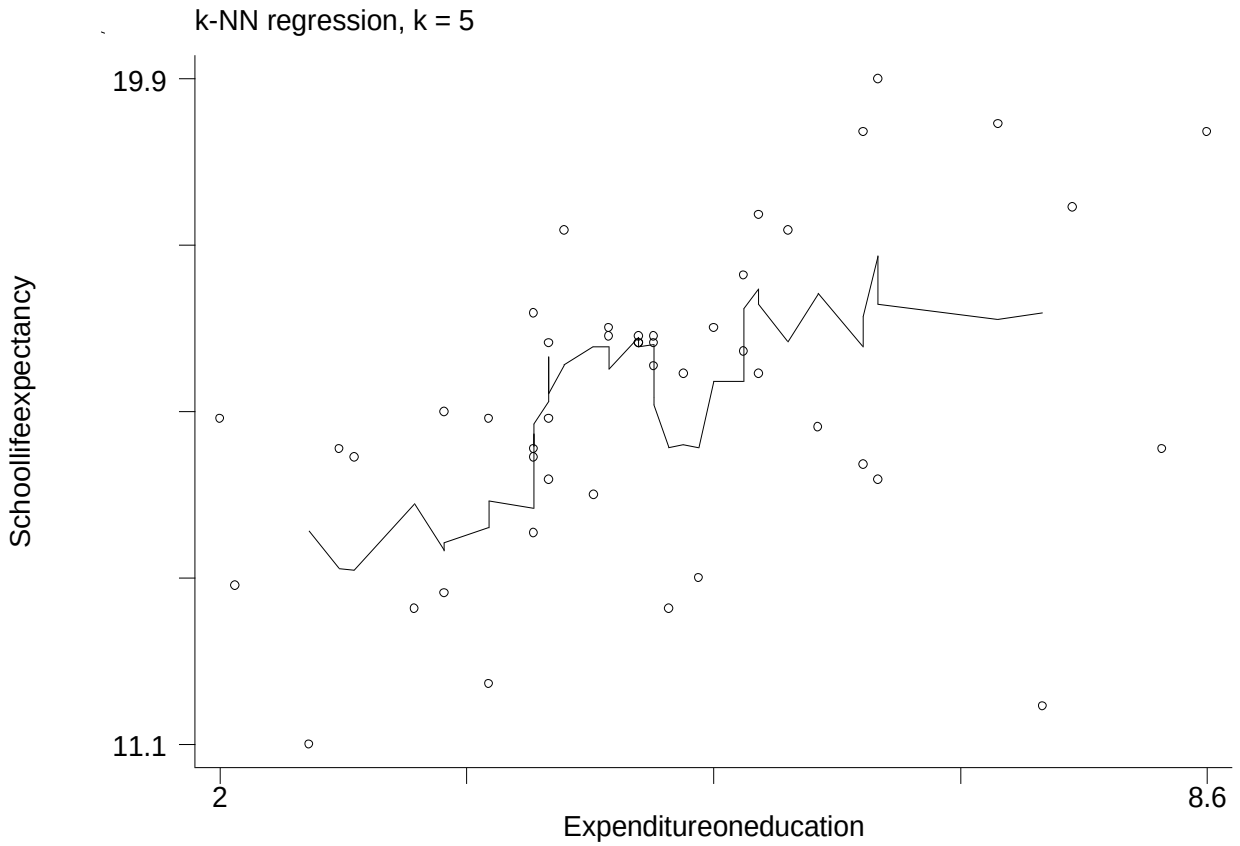
		Expenditureoneducation	Governmentexpenditureoneducat	Schoollifeexpectancy	Assessmentinreadingmathemati	Pupilteacherratiosecondary	Tertiaryenrolment	Graduatesinscienceandenginee	Tertiaryinboundmobility
Pearson Correlation	Expenditureoneducation	1.000	.770	.468	.470	-.123	.348	.046	.063
	Governmentexpenditureoneducat	.770	1.000	.421	.519	-.253	.379	-.004	.050
	Schoollifeexpectancy	.468	.421	1.000	.780	-.201	.802	-.205	-.014
	Assessmentinreadingmathemati	.470	.519	.780	1.000	-.459	.694	-.008	.141
	Pupilteacherratiosecondary	-.123	-.253	-.201	-.459	1.000	-.086	.055	-.283
	Tertiaryenrolment	.348	.379	.802	.694	-.086	1.000	-.081	-.271
	Graduatesinscienceandenginee	.046	-.004	-.205	-.008	.055	-.081	1.000	-.027
	Tertiaryinboundmobility	.063	.050	-.014	.141	-.283	-.271	-.027	1.000
Sig. (1-tailed)	Expenditureoneducation	.	.000	.001	.001	.222	.013	.388	.349
	Governmentexpenditureoneducat	.000	.	.003	.000	.055	.007	.491	.379
	Schoollifeexpectancy	.001	.003	.	.000	.104	.000	.100	.465
	Assessmentinreadingmathemati	.001	.000	.000	.	.001	.000	.480	.190
	Pupilteacherratiosecondary	.222	.055	.104	.001	.	.296	.365	.037
	Tertiaryenrolment	.013	.007	.000	.000	.296	.	.307	.044
	Graduatesinscienceandenginee	.388	.491	.100	.480	.365	.307	.	.435
	Tertiaryinboundmobility	.349	.379	.465	.190	.037	.044	.435	.

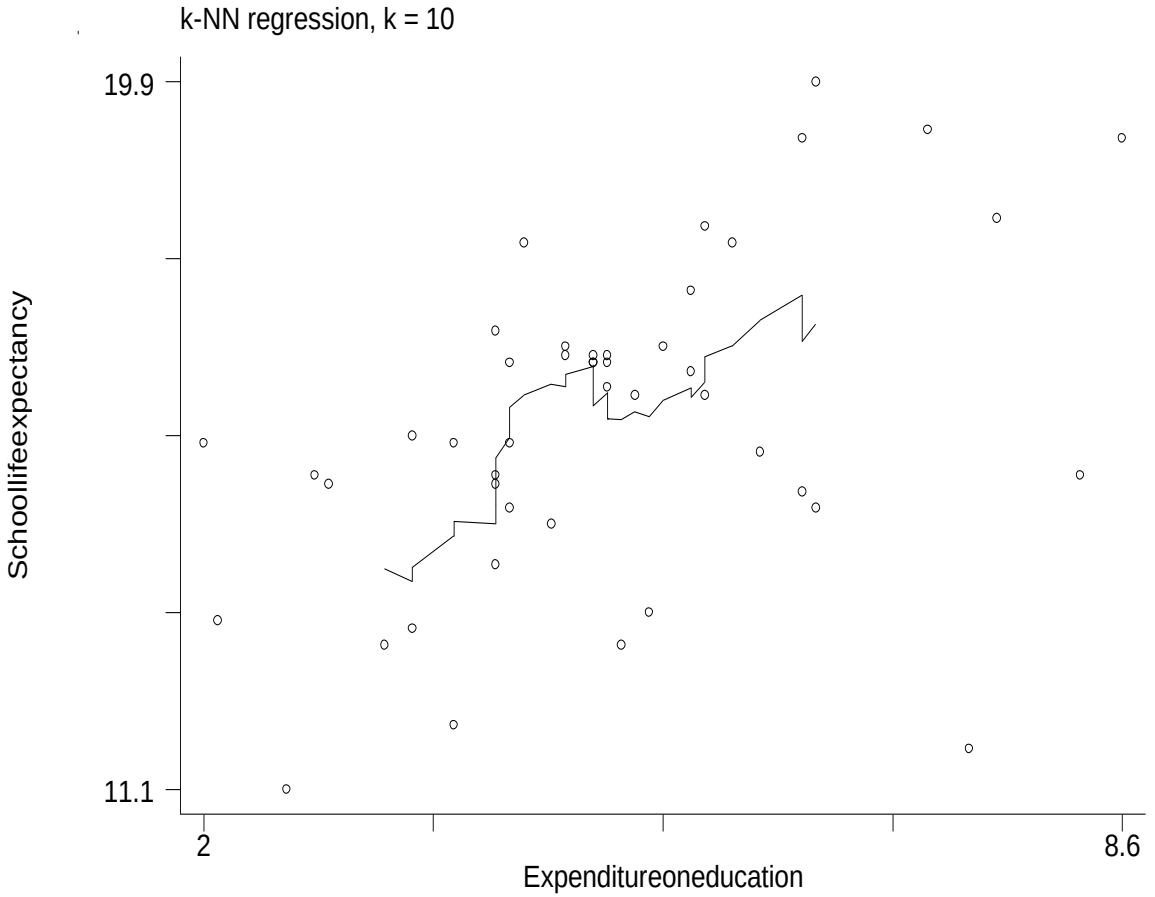
KNN-K(1,3) techniques –STATA output with one predictor





KNN-K(5, 10) techniques –STATA output with one predictor





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