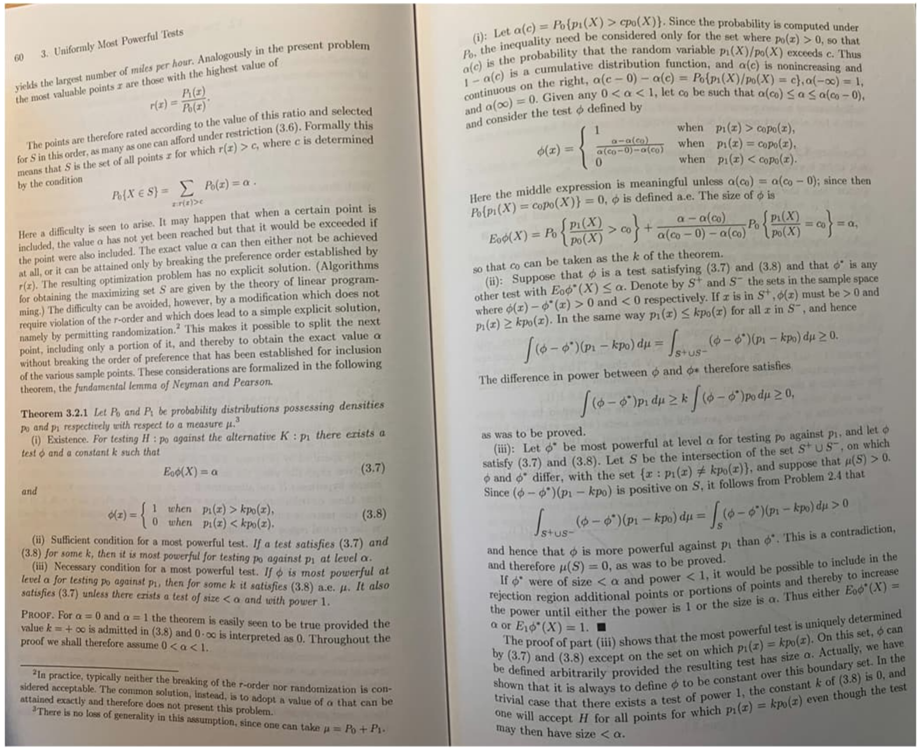
Statistisches Bundesamt

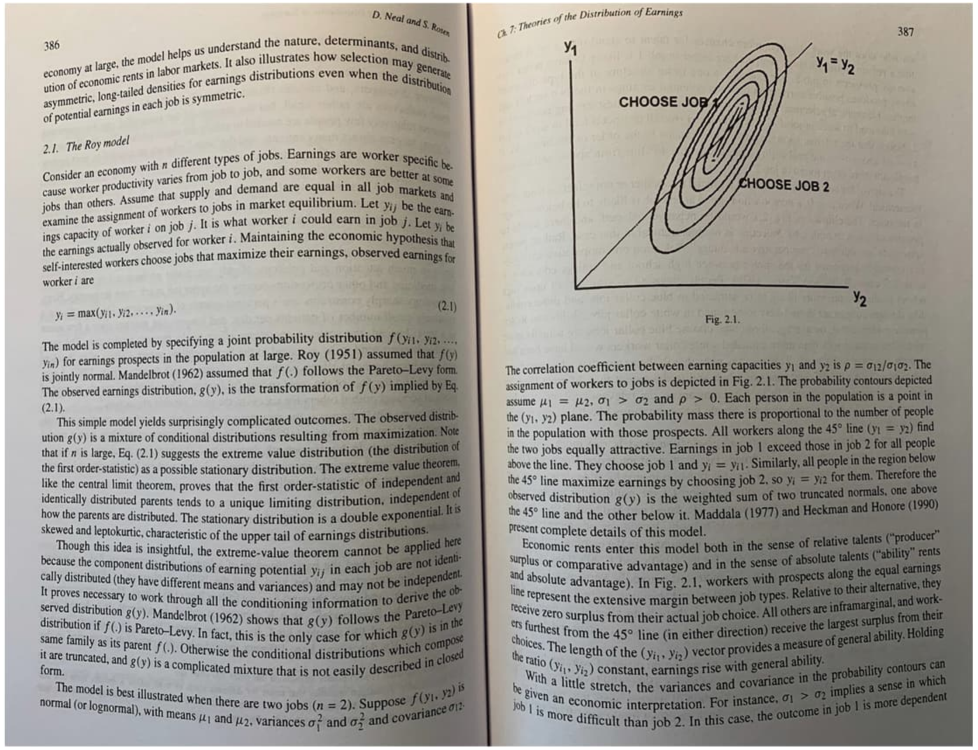
Reflections regarding the procedure and methodology of statistical data editing

October, 2025, uOttawa

Federal official statistics in Germany

It is not about …





Lehmann EL, Romano JP (2005) Testing Statistical Hypotheses, 3rd edition, Springer ; Atkinson AB, Bourguignon F (2000) Handbook of Income Distribution Vol. 1, North Holland

Federal official statistics in Germany

In is about ...

Die Statistik für Bundeszwecke (Bundesstatistik) hat im föderativ gegliederten Gesamtsystem der amtlichen Statistik die Aufgabe, laufend Daten über Massenerscheinungen zu erheben, zu sammeln, aufzubereiten, darzustellen und zu analysieren. Für sie gelten die Grundsätze der Neutralität, Objektivität und fachlichen Unabhängigkeit. Sie gewinnt die Daten unter Verwendung wissenschaftlicher Erkenntnisse und unter Einsatz der jeweils sachgerechten Methoden und Informationstechniken. Durch die Ergebnisse der Bundesstatistik werden gesellschaftliche, wirtschaftliche und ökologische Zusammenhänge für Bund, Länder einschließlich Gemeinden und Gemeindeverbände, Gesellschaft, Wirtschaft, Wissenschaft und Forschung aufgeschlüsselt. Die Bundesstatistik ist Voraussetzung für eine am Sozialstaatsprinzip ausgerichtete Politik. [...]

In the federally structured overall system of official statistics, statistics for federal purposes (federal statistics) have the task of continuously collecting, collating, processing, presenting and analysing data on mass phenomena. It is governed by the principles of neutrality, objectivity and professional independence. It collects data using scientific knowledge and appropriate methods and information technology. The results of federal statistics provide a breakdown of social, economic and ecological relationships for the Federation, the Länder including municipalities and municipal associations, society, the economy, science and research. Federal statistics are a prerequisite for a policy oriented towards the welfare state principle. [...]

§ 1 BStatG and an unauthorised translation

Federal official statistics in Germany

In a nutshell ...

- Europe in figures
- Countries & regions
- Government
- Labour
- Economy
- Economic sectors & enterprises
- Society & environment

400 sets of statistics

thereof

surveys	calculations	registers
323	71	6

182 primary surveys (57 %) 141 secondary surveys (43 %)

153 centralised statistics (38 %) 247 decentralised statistics (62 %)

as of February 2025

Federal official statistics in Germany

In a nutshell …

- Germany‘s federal structure – regional decentralisation
- The 14 statistical offices of the Länder are not subject to directives from the Federal Statistical Office
- Division of labour between federal and Länder level. Destatis is responsible for:
  - Methodological and technical preparation
  - Coordination of statistics production
  - Standardisation
  - Compilation / dissemination of the federal result
  - Data collection (centralised surveys)
  - International representation of the German statistical system

Federal official statistics in Germany

In a nutshell ...

- Methodological and technical preparation
  - whenever necessary for producing federal statistics in a uniform and high quality manner
  - also includes methodological decisions on e.g. sampling, classification, editing and imputation, weighting
- Not all methodological questions have already been solved
  - Need for exchange with other NSIs and with academia
  - However, there is no institutionalised collaboration with academia (besides EMOS)
    - No standardised exchange between Destatis and universities
    - Often difficult to find time to work on a scientific question

Federal official statistics in Germany

How we work: The Generic Statistical Business Process Model (GSBPM)

Specify needs	Design	Build	Collect	Process	Analyse	Disseminate	Evaluate
1.1 Identify needs	2.1 Design outputs	3.1 Reuse or build collection instruments	4.1 Create frame and select sample	5.1 Integrate data	6.1 Prepare draft outputs	7.1 Update output systems	8.1 Gather evaluation inputs
1.2 Consult and confirm needs	2.2 Design variable descriptions	3.2 Reuse or build processing and analysis components	4.2 Set up collection	5.2 Classify and code	6.2 Validate outputs	7.2 Produce dissemination products	8.2 Conduct evaluation
1.3 Establish output objectives	2.3 Design collection	3.3 Reuse or build dissemination components	4.3 Run collection	5.3 Review and validate	6.3 Interpret and explain outputs	7.3 Manage release of dissemination products	8.3 Agree an action plan
1.4 Identify concepts	2.4 Design frame and sample	3.4 Configure workflows	4.4 Finalise collection	5.4 Edit and impute	6.4 Apply disclosure control	7.4 Promote dissemination products	
1.5 Check data availability	2.5 Design processing and analysis	3.5 Test production systems		5.5 Derive new variables and units	6.5 Finalise outputs	7.5 Manage user support	
1.6 Prepare and submit business case	2.6 Design production systems and workflow	3.6 Test statistical business process		5.6 Calculate weights			
		3.7 Finalise production systems		5.7 Calculate aggregates			
				5.8 Finalise data files			

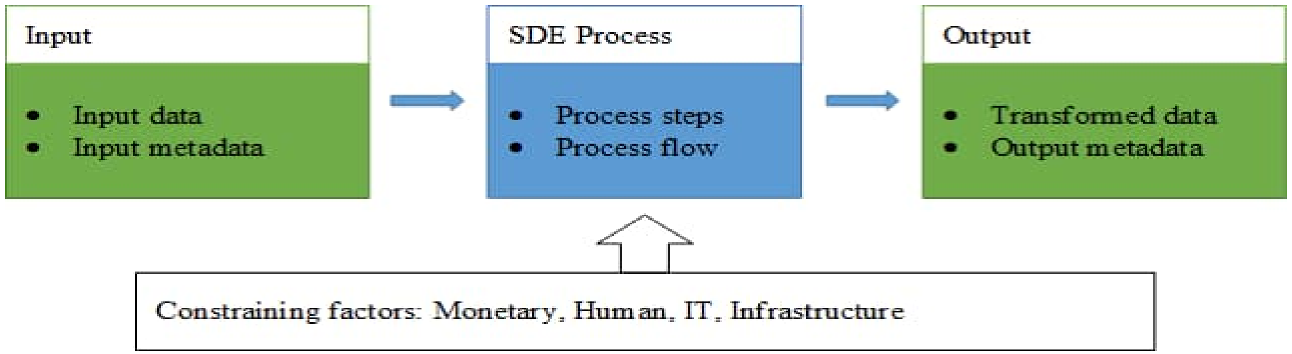
https://statswiki.unece.org/display/GSBPM

Editing and imputation

Basics

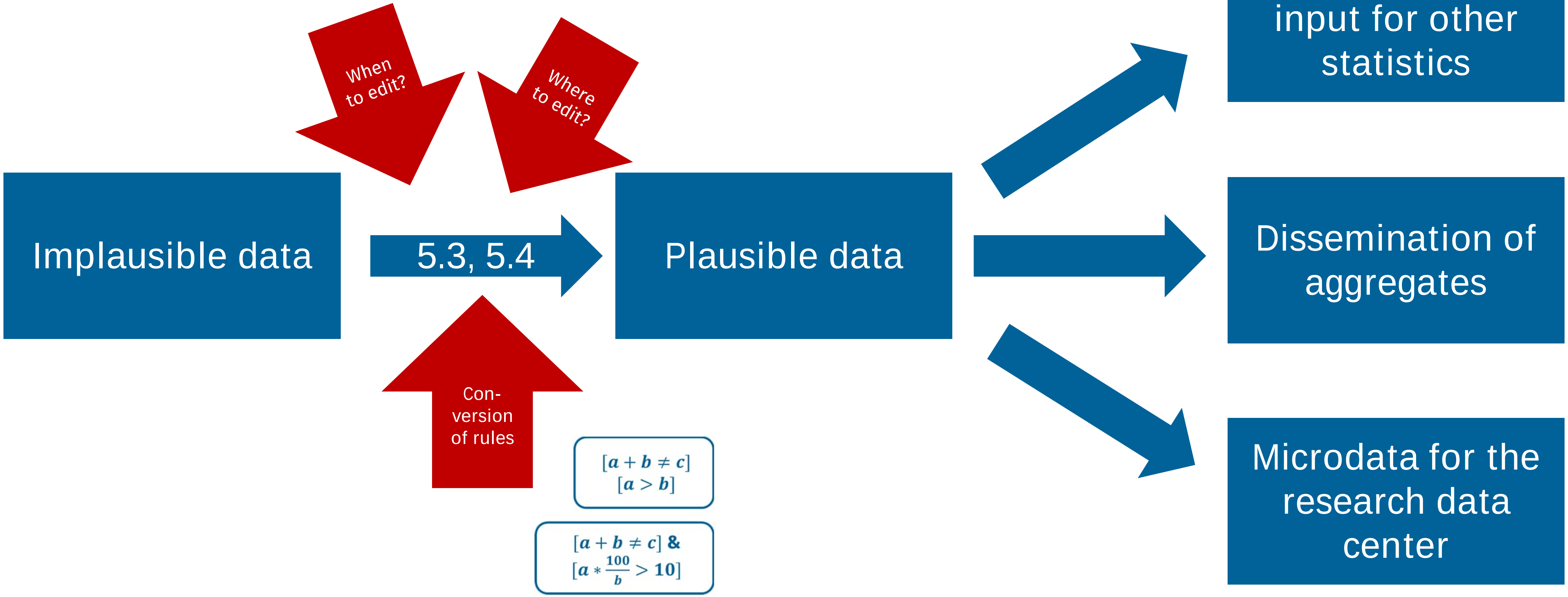
> “The statistical data editing (SDE) process can be represented as follows [...]: data and metadata are provided as an input, a series of activities are performed to assess data plausibility, identify potential problems and remedy the problems; and transformed data are produced as an output. The process is set according to constraining factors as shown.”

Generic Statistical Data Editing Model (GSDEM) version 2.0, https://unece.org/statistics/documents/2019/06/gsdem-v20 · Also very worth reading: Scholtus S (2025) The Unknown Future of Statistical Data Editing: Some Imputations. Journal of Official Statistics, 41(3), 901–911 · And very famous: Hidiroglou MA, Berthelot J-M (1986) Statistical Editing and Imputation for Periodic Business Surveys. Survey Methodology, 12(1), 73–83



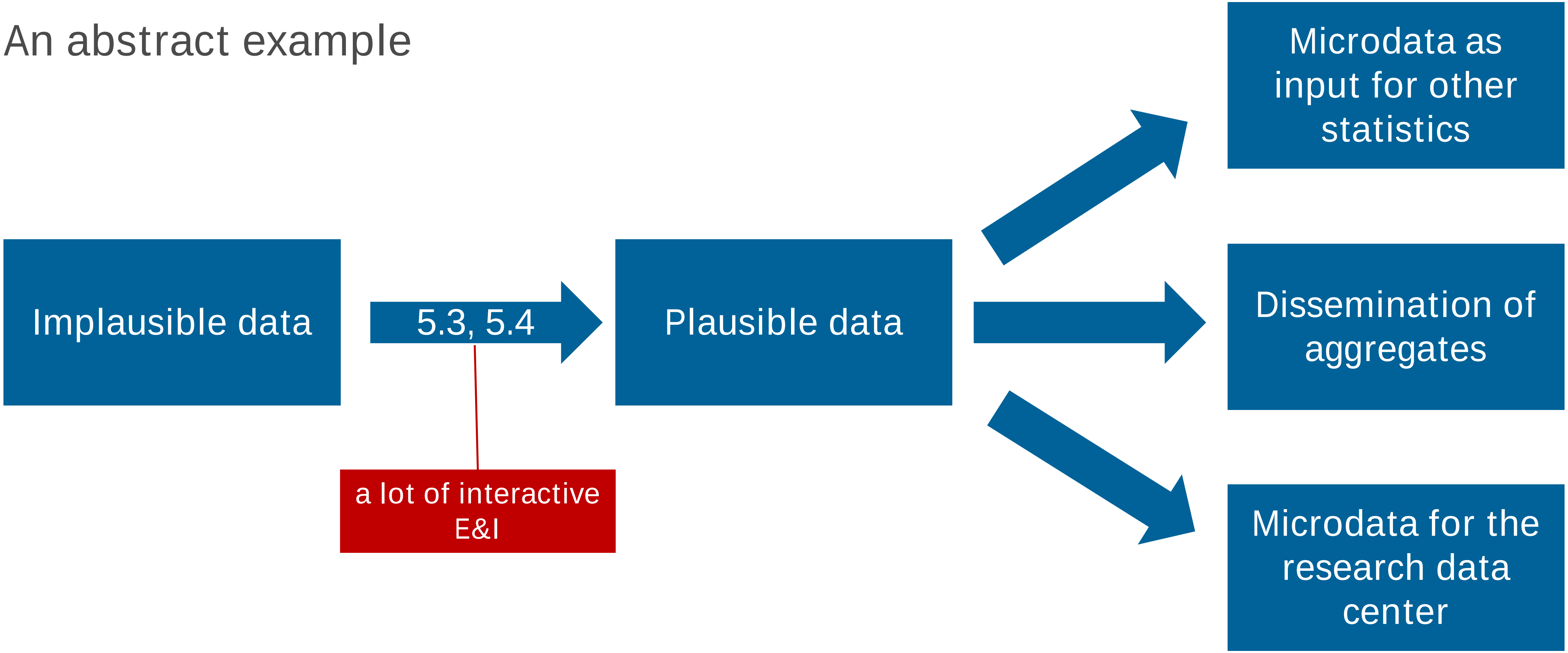
Editing and imputation

An abstract example



Editing and imputation

An abstract example



Editing and imputation

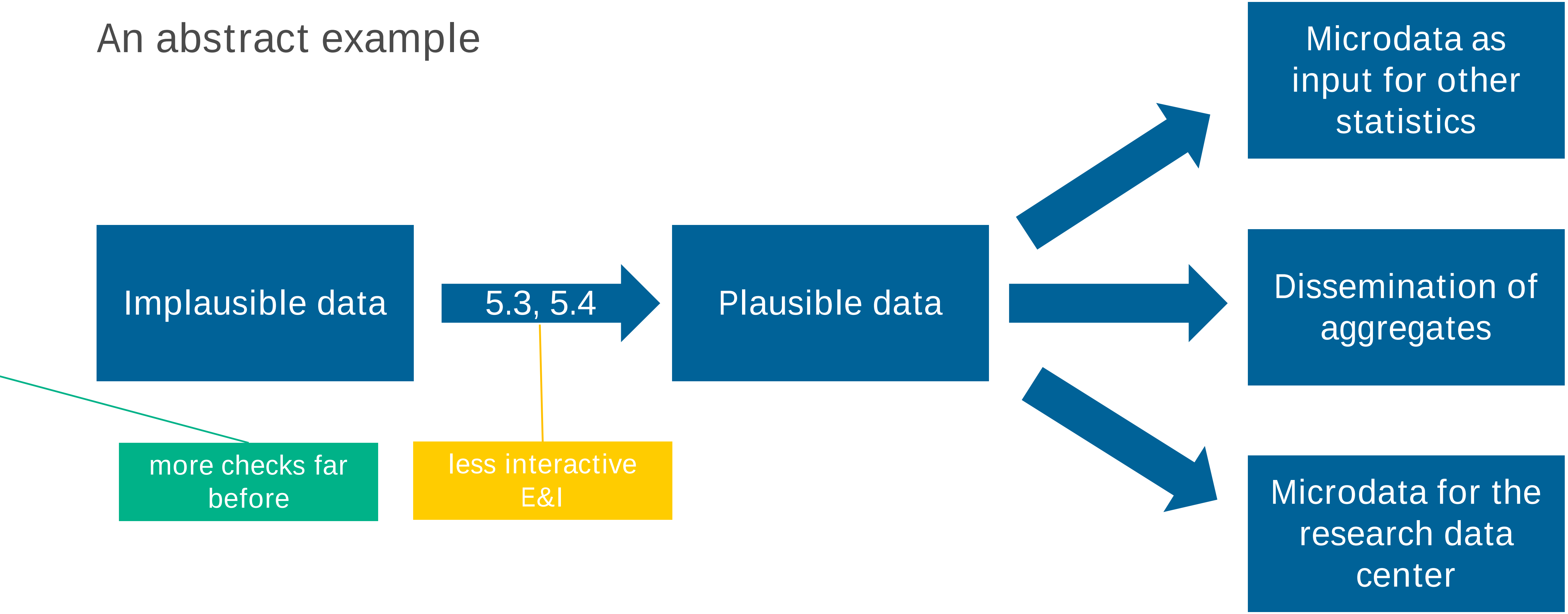
An abstract example – The effort point of view

"The occurrence of nonresponse and, especially, errors in the observed data makes it necessary to carry out an extensive process of checking the collected data, and, when necessary, correcting them. This checking and correction process is referred to as statistical data editing and imputation. [...] Any improvement in the efficiency of the editing and imputation process should [...] be highly welcomed by NSIs."

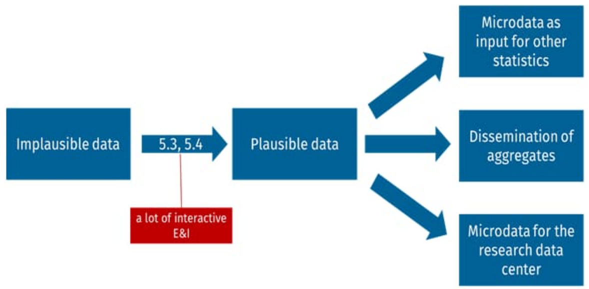
de Waal T, Pannekoek J, Scholtus S (2010) Handbook of Statistical Data Editing and Imputation. Wiley.

Editing and imputation

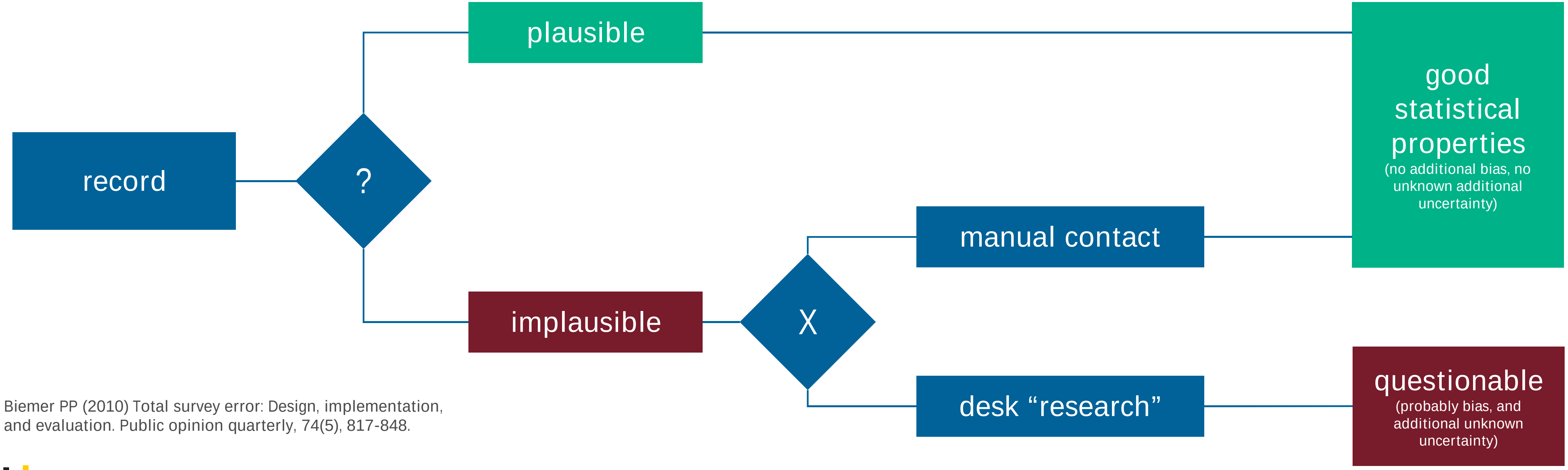
An abstract example



Editing and imputation

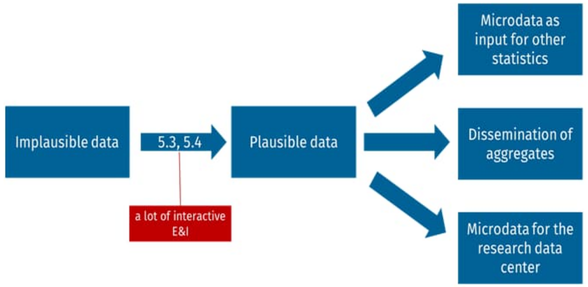


An abstract example – the statistical point of view

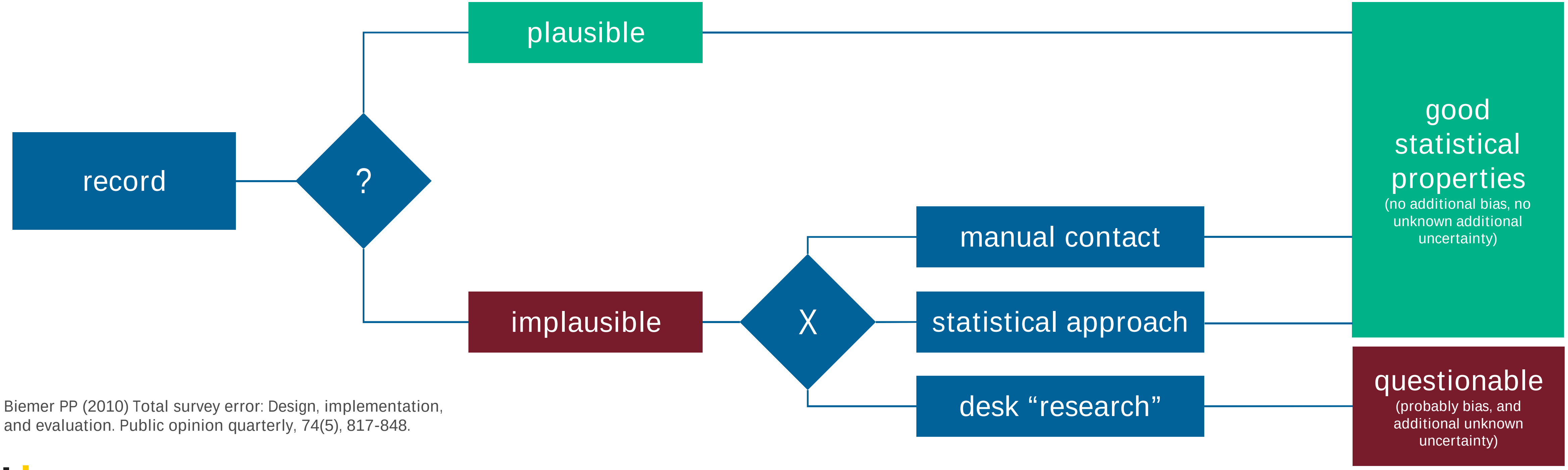


Biemer PP (2010) Total survey error: Design, implementation, and evaluation. Public opinion quarterly, 74(5), 817-848.

Editing and imputation

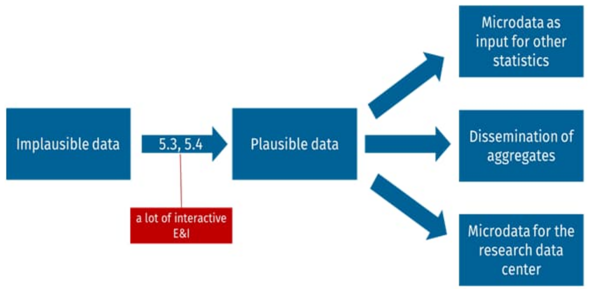


An abstract example – the statistical point of view

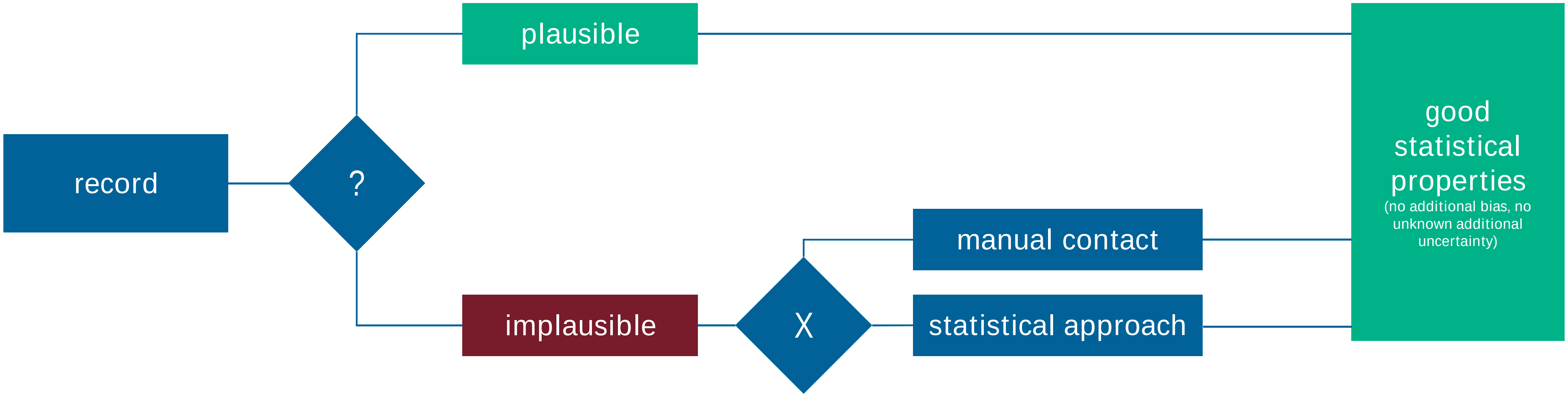


Biemer PP (2010) Total survey error: Design, implementation, and evaluation. Public opinion quarterly, 74(5), 817-848.

Editing and imputation



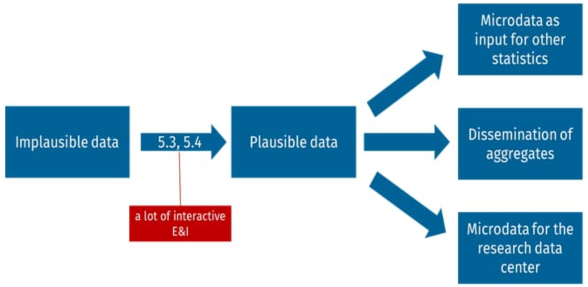
An abstract example – the statistical point of view

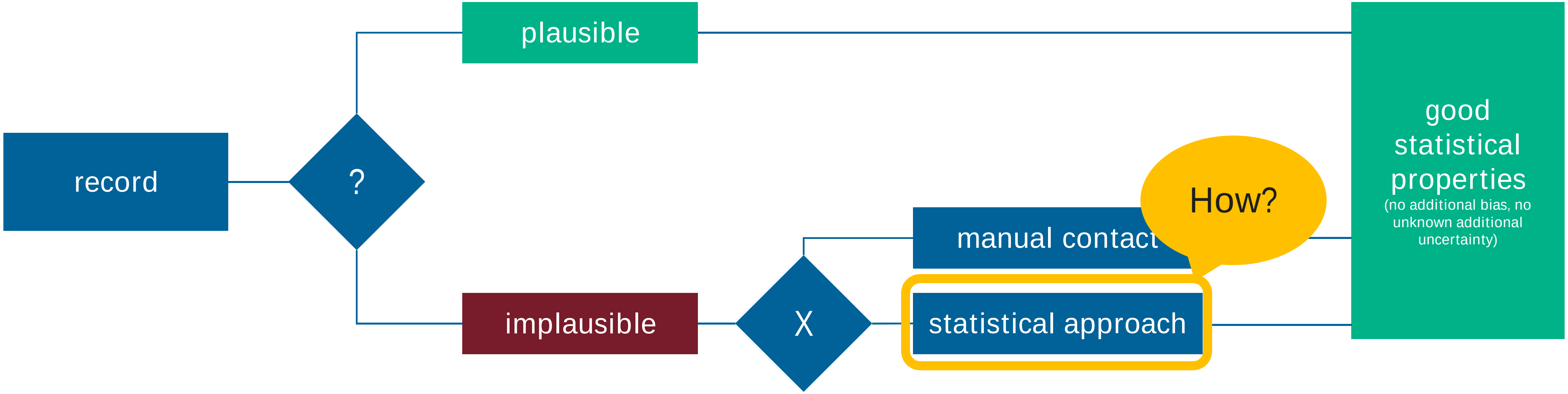


Biemer PP (2010) Total survey error: Design, implementation, and evaluation. Public opinion quarterly, 74(5), 817-848.

Editing and imputation

An abstract example – the statistical point of view





Biemer PP (2010) Total survey error: Design, implementation, and evaluation. Public opinion quarterly, 74(5), 817-848.

Editing and imputation

Aim: The data has to be correct …

- What do you mean by correct?
  - Consistency of the data? (According to what?)
  - Is some uncertainty allowed?
  - Is a certain amount of fuzziness allowed / needed?
    (input for other statistics, aggregates, research data center)

- What exactly has to be correct?
  - Every entry of the data set? Or something else? Of which data set:
    - » The one for the research data center?
    - » The one to be submitted to Eurostat?
    - » The one used for further internal production?

Editing and imputation

Aim: The data has to be correct …

- How to measure that?
  - “Only entries that have been checked by an specialised employee are correct …”
  - “Let’s call the enterprise and ask for the correct value.”
  - “They often make mistakes here …”

- How much may we change the submitted values?
- When do we cross the border from statistical editing to (possibly unintentional) forgery?

Editing and imputation

Statistical answers to the question of correctness and accuracy

ImpAct (Imputation Assessment and Comparison Tool)

- Graphical univariate distribution analysis
  - kernel density curves
  - histograms
  - box-plots
  - ...
- But: assessing distributional aspects of imputation approaches only based on visual analysis may be misleading!

Gray D (2019) A Generalized Framework to Evaluate Imputation Strategies: Recent Developments. In: JSM Proceedings, Government Statistics Section. American Statistical Association. 1861–1870.

Editing and imputation

Statistical answers to the question of correctness and accuracy

1. Predictive accuracy
2. Ranking accuracy
3. Estimation accuracy
4. Distributional accuracy
5. Imputation plausibility

Chambers R (2006) Evaluation Criteria for Editing and Imputation in Euredit. In: Statistical Data Editing. vol. 3. United Nations Statistical Commission and United Nations Economic Commission for Europe.

Editing and imputation

Statistical answers to the question of correctness and accuracy

1. Predictive accuracy:

The imputation procedure should maximise preservation of true values. That is, it should result in imputed values that are "close" as possible to the true values.

("reproduction of the true values")

Chambers R (2006) Evaluation Criteria for Editing and Imputation in Euredit. In: Statistical Data Editing. vol. 3. United Nations Statistical Commission and United Nations Economic Commission for Europe.

Editing and imputation

Basics

"Imputation is a method for the analysis of data with missing values, where missing values are replaced by estimates and the filled-in data are analyzed by complete-data methods. [...] In fact, the main reason for imputation is not to recover the information in the missing values, which is lost and usually not recoverable, but rather to allow the information in observed values in the incomplete cases to be retained."

Little RJ (2011) Imputation. In: Lovric M (2011) International Encyclopedia of Statistical Science, Springer.

Editing and imputation

Statistical answers to the question of correctness and accuracy

2. Ranking accuracy:

The imputation procedure should maximise preservation of order in the imputed values. That is, it should result in ordering relationships between imputed values that are the same (or very similar) to those that hold in the true values.

Chambers R (2006) Evaluation Criteria for Editing and Imputation in Euredit. In: Statistical Data Editing. vol. 3. United Nations Statistical Commission and United Nations Economic Commission for Europe.

Editing and imputation

Statistical answers to the question of correctness and accuracy

3. Estimation accuracy: (also inferential accuracy)

The imputation procedure should reproduce the lower order moments of the distributions of the true values. In particular, it should lead to unbiased and efficient inferences for parameters of the distribution of the true values (given that these true values are unavailable).

Imputation is a method for the analysis of data with missing values, where missing values are replaced by estimates and the filled-in data are analyzed by complete-data methods ... (see Little some slides before)

Chambers R (2006) Evaluation Criteria for Editing and Imputation in Euredit. In: Statistical Data Editing. vol. 3. United Nations Statistical Commission and United Nations Economic Commission for Europe.

Editing and imputation

Statistical answers to the question of correctness and accuracy

4. Distributional accuracy:

The imputation procedure should preserve the distribution of the true data values. That is, marginal and higher order distributions of the imputed data values should be essentially the same as the corresponding distributions of the true values.

Imputation is a method for the analysis of data with missing values, where missing values are replaced by estimates and the filled-in data are analyzed by complete-data methods … (see Little some slides before)

Chambers R (2006) Evaluation Criteria for Editing and Imputation in Euredit. In: Statistical Data Editing. vol. 3. United Nations Statistical Commission and United Nations Economic Commission for Europe.

Editing and imputation

Statistical answers to the question of correctness and accuracy

5. Imputation plausibility:

The imputation procedure should lead to imputed values that are plausible. In particular, they should be acceptable values as far as the editing procedure is concerned.

Chambers R (2006) Evaluation Criteria for Editing and Imputation in Euredit. In: Statistical Data Editing. vol. 3. United Nations Statistical Commission and United Nations Economic Commission for Europe.

Editing and imputation

How to measure these accuracies?

1. Predictive accuracy → For every entry something like dist(x_{imp}, x_{true}).
2. Ranking accuracy → Counting (and weighting) violations of the order.
3. Estimation accuracy → For every lower (mixed) moment s.th. like dist$(\widehat{\theta_{imp}}, \widehat{\theta_{true}})$.
4. Distributional accuracy → Something like dist(F_{imp}, F_{true}).
5. Imputation plausibility → Check against edit rules.

Chambers R (2006) Evaluation Criteria for Editing and Imputation in Euredit. In: Statistical Data Editing. vol. 3. United Nations Statistical Commission and United Nations Economic Commission for Europe.

References for the next slides:
Thurow M, Dumpert F, Ramosaj B, Pauly M (2021) Imputing missings in official statistics for general tasks – our vote for distributional accuracy. Statistical Journal of the IAOS, 37, 1379–1390 ·
Thurow M, Dumpert F, Ramosaj B, Pauly M (2021) Goodness (of fit) of Imputation Accuracy: The GoodImpact Analysis. https://doi.org/10.48550/arXiv.2101.07532 ·
Thurow M, Dumpert F, Ramosaj B, Pauly M (2024) Assessing the multivariate distributional accuracy of common imputation methods. Statistical Journal of the IAOS, 40, 99–108

Predictive accuracy measures

Proportion of falsely classified / imputed entries (PFC)

$$PFC = \frac{\sum_{j \in \boldsymbol{C}} \sum_{i=1}^{n} \left(m_{ij} \cdot 1\left\{ x_{ij}^{(imp)} \neq x_{ij}^{(true)} \right\} \right)}{\sum_{j \in \boldsymbol{C}} \sum_{i=1}^{n} m_{ij}}$$

$\boldsymbol{X} = \left(x_{ij}\right)_{i=1,\dots,n, j=1,\dots,d} \in \mathbb{R}^{n \times d}$ (n observations in d variables),
$\boldsymbol{M} = \left(m_{ij}\right)_{i=1,\dots,n, j=1,\dots,d} \in \{0,1\}^{n \times d}$ indicates whether an entry is missing or not,
$\boldsymbol{C} \subset \{1, \dots, d\}$ is the subset of categorical variables.

Predictive accuracy measures

Normalised root mean squared error (NRMSE)

$$NRMSE = \sqrt{\frac{\sum_{j\in \boldsymbol{N}} \sum_{i=1}^{n} \left(m_{ij} \cdot \left(x_{ij}^{(imp)} - x_{ij}^{(true)} \right)^2 \right)}{\sum_{j\in \boldsymbol{N}} \sum_{i=1}^{n} \left(m_{ij} \cdot \left(x_{ij}^{(imp)} - \bar{x}^{(true)} \right)^2 \right)}}$$

$\boldsymbol{X} = \left(x_{ij}\right)_{i=1,\dots,n,j=1,\dots,d} \in \mathbb{R}^{n\times d}$ (n observations in d variables),
$\boldsymbol{M} = \left(m_{ij}\right)_{i=1,\dots,n,j=1,\dots,d} \in \{0,1\}^{n\times d}$ indicates whether an entry is missing or not,
$\boldsymbol{N} \subset \{1,\dots,d\}$ is the subset of metric variables.

Distributional accuracy measures (univariate)

Cramér’s V for nominal variables j

$$\chi_j^2 = \sum_{\substack{\text{entries of the} \\ \text{contingency table of variable } j}} \frac{(O - E)^2}{E}$$

$$V_j = \sqrt{\frac{\chi_j^2 / n}{\#\text{categories of variable } j - 1}}$$

Distributional accuracy measures (univariate)

(Two sample) Kolmogorov-Smirnov-statistic (KS) for metric and ordinal variables j

$$k_j^{(0)} = \max_{z \in \mathcal{T}_j} \left| F_j^{(true)}(z) - F_j^{(imp)}(z) \right|$$

F_j are the empirical distribution functions of variable j in the original and the imputed data set, respectively,
$\mathcal{T}_j$ is the support of variable j.

Distributional accuracy measures (univariate)

Kolmogorov-Smirnov test

$$H_0 : F_j^{(true)} = F_j^{(imp)} \quad vs. \quad H_1 : F_j^{(true)} \neq F_j^{(imp)}$$

- We calculated a permutation-based p-value (asymptotically exact), based on the definition of the p-value
- p-value = probability of obtaining test results at least as extreme as the result actually observed, under the assumption that the null hypothesis is correct
- Three steps (for every variable j separately, average if multiply imputed) needed

Distributional accuracy measures (univariate)

Kolmogorov-Smirnov test

$$H_0 : F_j^{(true)} = F_j^{(imp)} \quad vs. \quad H_1 : F_j^{(true)} \neq F_j^{(imp)}$$

- Three steps (for every variable j separately, average if multiply imputed) needed

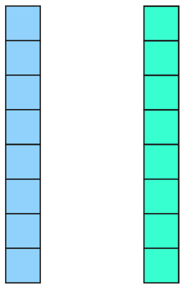
1. Compute the actual statistics $k_j^{(0)}$ $\qquad k_j^{(0)} = \max_{z \in \mathcal{T}_j} \left| F_j^{(true)}(z) - F_j^{(imp)}(z) \right|$
2. Permute the observations of variable j of the original data set and the imputed data set(s) – and compute (if multiply: including averaging in the end) $k_j^{(l)}$
3. Repeat Step 2 $\#perm$ times → $k_j^{(l)}, l = 0, \ldots, \#perm$

Distributional accuracy measures (univariate)

Kolmogorov-Smirnov test

$$H_0 : F_j^{(true)} = F_j^{(imp)} \quad vs. \quad H_1 : F_j^{(true)} \neq F_j^{(imp)}$$

- Step 2: Permute the observations of variable j of the original data set and the imputed data set



true

actually imputed

and $H_0 : F_j^{(true)} = F_j^{(imp)}$ leads to

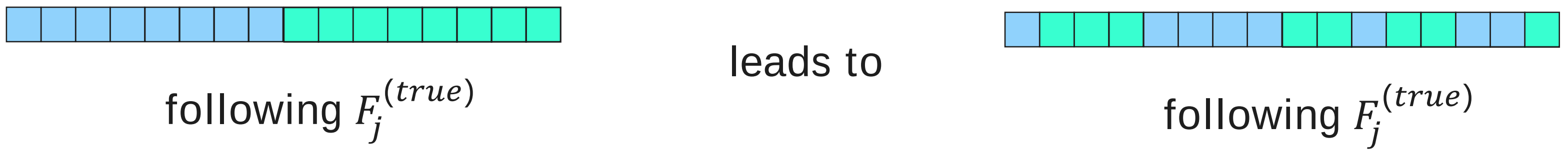
following $F_j^{(true)}$

Distributional accuracy measures (univariate)

Kolmogorov-Smirnov test

$$H_0 : F_j^{(true)} = F_j^{(imp)} \quad vs. \quad H_1 : F_j^{(true)} \neq F_j^{(imp)}$$

- Step 2: Permute the observations of variable j of the original data set and the imputed data set



Distributional accuracy measures (univariate)

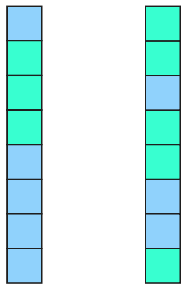
Kolmogorov-Smirnov test

$$H_0 : F_j^{(true)} = F_j^{(imp)} \quad vs. \quad H_1 : F_j^{(true)} \neq F_j^{(imp)}$$

- Step 2: Permute the observations of variable j of the original data set and the imputed data set

following $F_j^{(true)}$

leads to the new possible comparison



$\rightarrow k_j^{(l)}$

Distributional accuracy measures (univariate)

Kolmogorov-Smirnov test

$$H_0 : F_j^{(true)} = F_j^{(imp)} \quad vs. \quad H_1 : F_j^{(true)} \neq F_j^{(imp)}$$

- p-value = probability of obtaining test results at least as extreme as the result actually observed, under the assumption that the null hypothesis is correct
- → p-value = $\frac{1}{\#perm + 1} \left(\sum_{l=0}^{\#perm} 1 \left\{ k_j^{(l)} \geq k_j^{(0)} \right\} \right)$
- (If all possible permutations could be evaluated the test would be exact.)

Laplace's definition of probability: The probability of an event is the ratio of the number of cases favorable to it, to the number of all cases possible.

Accuracy measures (univariate)

	categorical		
	nominal	ordinal	metric
predictive accuracy	Proportion of falsely classified / imputed entries (PFC)		NRMSE
distributional accuracy	Cramér's V	Kolmogorov-Smirnov test	

Estimation accuracy measures (univariate)

Some important univariate values

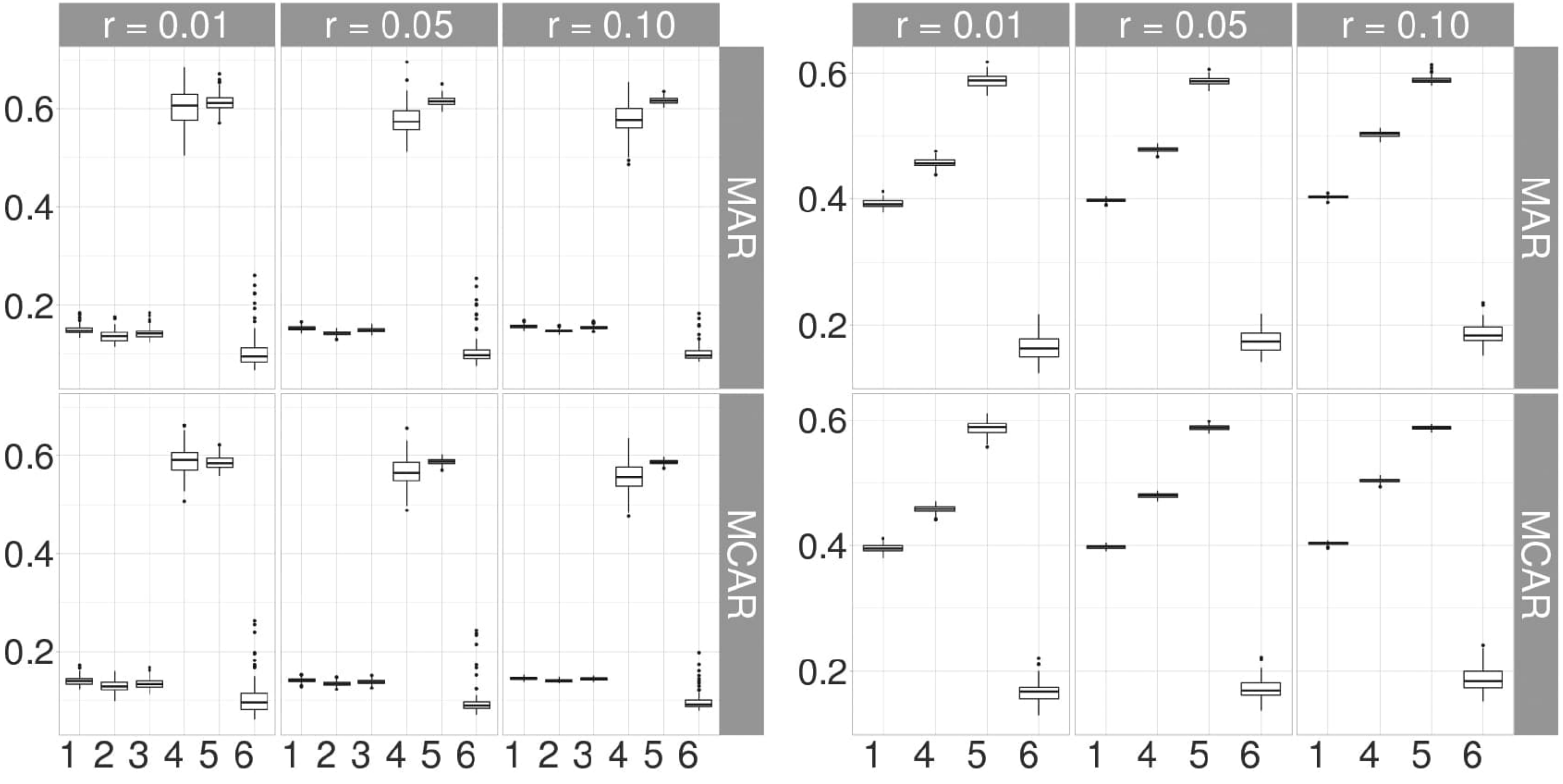
- First four moments of key variables
- Important quantiles

→ Squared difference between true and imputed results, i.e.
dist$(\widehat{\theta_{imp}}, \widehat{\theta_{true}}) = (\widehat{\theta_{imp}} - \widehat{\theta_{true}})^2$

From a theoretical (!) point of view:

If the distributions of true and imputed values coincide, also the moments and the quantiles should coincide.

Some results on predictive accuracy



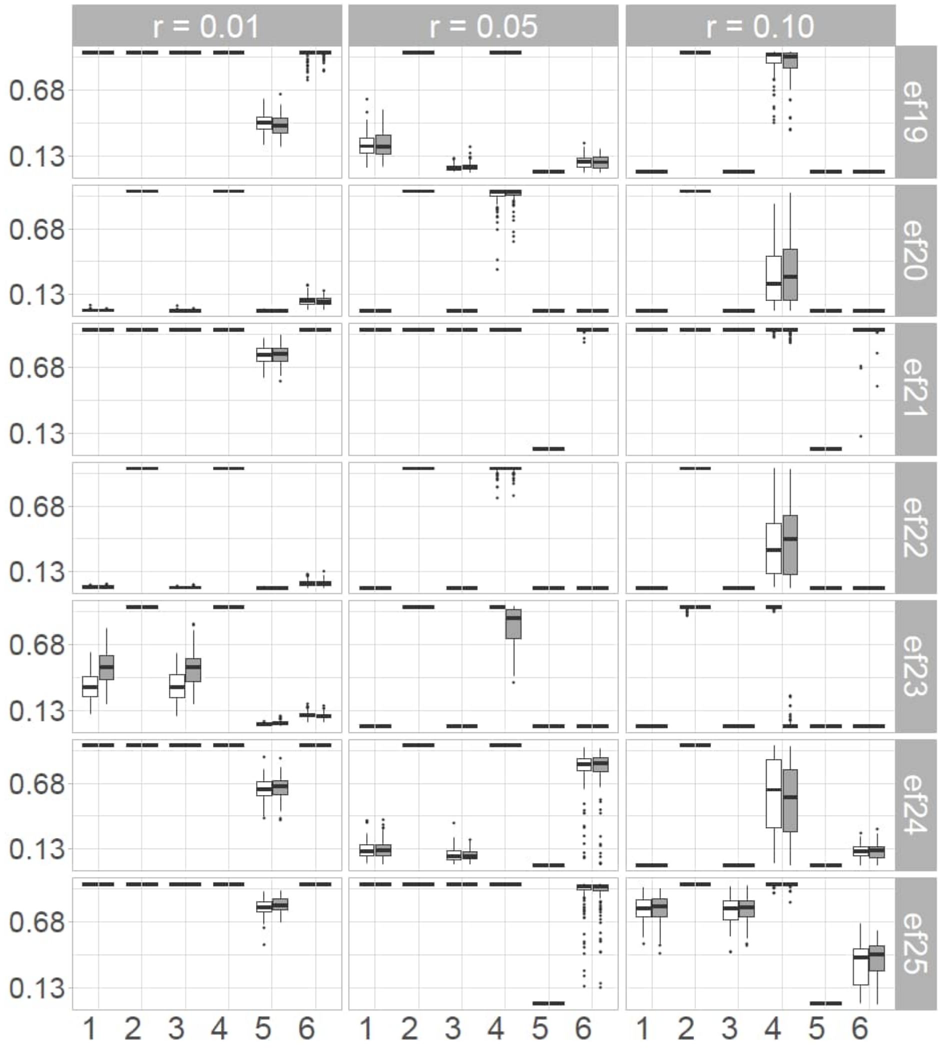
(a) NRMSE (b) PFC

MC = 100

1: Amelia
2: Mice.Pmm
3: Mice.Norm
4: Mice.RF
5: Naive
6: missRanger

The lower NRMSE and PFC the better.

Some results on distributional accuracy



MC = 100

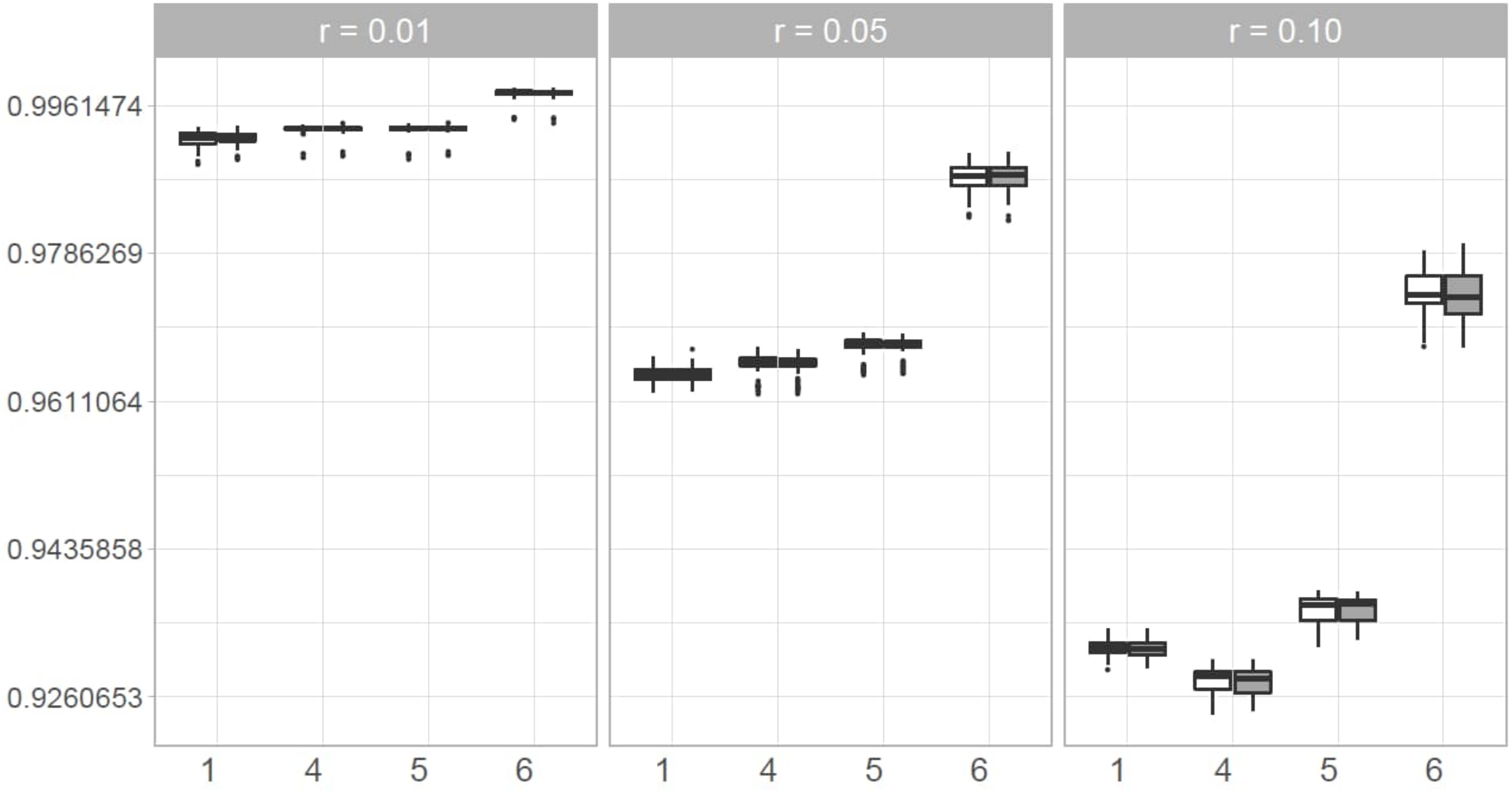
p-values of KS for some metric and ordinal variables

Left boxplot for MCAR, right one for MAR

1: Amelia
2: Mice.Pmm
3: Mice.Norm
4: Mice.RF
5: Naive
6: missRanger

The higher the p-value the less suspicious the difference (i.e., more or less, the better)

Some results on distributional accuracy



MC = 100

Boxplots of Cramér's V, averaged over all nominal variables

Left boxplot for MCAR, right one for MAR

1: Amelia
4: Mice.RF
5: Naive
6: missRanger

The higher Cramér's V the higher (i.e., more or less, the better) the association between original and imputed data.

Some results on estimation accuracy

Method	r	Mean 139.83	sd 47.66	sk. −1.29	kurt. 3.25	$q_{0.25}$ 120.00	$q_{0.5}$ 165.00	$q_{0.75}$ 174.00
Amelia	0.01	0	0	0	0	0.22	–	–
	0.05	0	0	0	0	0.88	–	–
	0.10	0	0	0	0	1.00	–	–
Mice.Pmm	0.01	0	0	0	0	0.15	–	–
	0.05	0	0	0	0	0.27	–	–
	0.10	0	0	0	0	0.24	–	–
Mice.Norm	0.01	0	0	0	0	0.71	–	–
	0.05	0	0	0	0	1.04	–	–
	0.10	0.10	0.01	0	0	2.51	–	–
Mice.RF	0.01	0	0	0	0	0.01	–	–
	0.05	0.02	0	0	0	0.10	–	–
	0.10	0.10	0.01	0	0	2.24	–	–
Naive	0.01	0	0.06	0	0	2.35	–	–
	0.05	0	1.46	0	0.03	41.57	–	–
	0.10	0.01	5.96	0	0.13	100	4	1
missRanger	0.01	0	0	0	0	0.02	–	–
	0.05	0	0	0	0	0.12	–	–
	0.10	0.01	0.14	0	0	0.27	0.03	0.10

paid hours

Mean squared deviation, i.e. dist($\widehat{\theta_{imp}}, \widehat{\theta_{true}}$) = $\left(\widehat{\theta_{imp}} - \widehat{\theta_{true}}\right)^2$, over the MC=100 Monte-Carlo iterations for MAR

"–" in the table denotes the value zero. A zero in the table means that the value is smaller than 0.01 but not zero.

The lower the deviation the better is the estimation accuracy for the quantile or the moment.

Some results on estimation accuracy

Method	r	Mean 2420.13	sd 1637.98	sk. 0.86	kurt. 3.41	$q_{0.25}$ 1202.00	$q_{0.5}$ 2175.00	$q_{0.75}$ 3283.00
Amelia	0.01	0	0	0	0	0.05	0.09	0.21
	0.05	0.04	0.05	0	0	0.33	0.52	1.10
	0.10	0.09	0.17	0	0	1.20	2.20	6.53
Mice.Pmm	0.01	0	0	0	0	–	0	0.01
	0.05	0.02	0.05	0	0	0.11	0.14	0.35
	0.10	0.09	0.55	0	0	0.80	0.55	1.41
Mice.Norm	0.01	0	0	0	0	0	0.01	0.04
	0.05	0.02	0.03	0	0	0.21	0.20	0.48
	0.10	0.09	0.41	0	0	1.79	0.60	1.36
Mice.RF	0.01	0.58	1.47	0	0	3.01	0.81	2.50
	0.05	12.56	37.65	0	0	62.64	6.46	77.04
	0.10	49.51	136.78	0	0	195.48	23.26	359.15
Naive	0.01	1.08	65.76	0	0	268.79	487.84	204.34
	0.05	5.79	1728.86	0	0.03	3659.23	11861.20	7328.44
	0.10	10.18	7120.55	0	0.14	15911.10	48965.05	29309.07
missRanger	0.01	0.02	1.00	0	0	1.56	0.13	0.77
	0.05	0.14	12.47	0	0	27.01	1.68	8.07
	0.10	0.42	27.50	0	0	49.29	4.51	22.43

gross monthly income

Mean squared deviation, i.e. dist$(\widehat{\theta_{imp}}, \widehat{\theta_{true}})$ = $(\widehat{\theta_{imp}} - \widehat{\theta_{true}})^2$, over the MC=100 Monte-Carlo iterations for MAR

"–" in the table denotes the value zero. A zero in the table means that the value is smaller than 0.01 but not zero.

The lower the deviation the better is the estimation accuracy for the quantile or the moment.

Distributional accuracy measures (multivariate)

Kolmogorov-Smirnov-based approach

- For a metric or ordinal variable we can calculate

$$k_j^{(0)} = \max_{z \in \mathcal{T}_j} \left| F_j^{(true)}(z) - F_j^{(imp)}(z) \right|$$

- For more than one variable there is no direct equivalent
- However, mathematical-statistics knows some things on joint distributions, e.g. the Cramér-Wold theorem

Distributional accuracy measures (multivariate)

Kolmogorov-Smirnov-based approach

Cramér-Wold theorem

Let $U = (U_1, \dots, U_s)^T$ and V = $(V_1, \dots, V_s)^T$ be random vectors in $\mathbb{R}^s$.

U and V follow the same distribution

if and only if

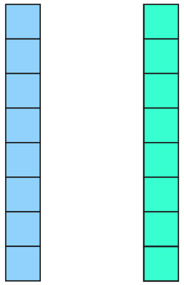
for every $t \in \mathbb{R}^s$ with $\|t\| = 1$: $t^T U$ follows the same distribution as $t^T V$.

Distributional accuracy measures (multivariate)

Kolmogorov-Smirnov-based approach

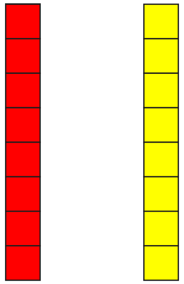
So far (univariate): just the marginals

KS for one specific variable j_1



true actually imputed

KS for one specific variable j_2



true actually imputed

$$k_{j_1}^{(0)} = \max_{z\in\mathcal{T}_{j_1}} \left|F_{j_1}^{(true)}(z) - F_{j_1}^{(imp)}(z)\right|$$

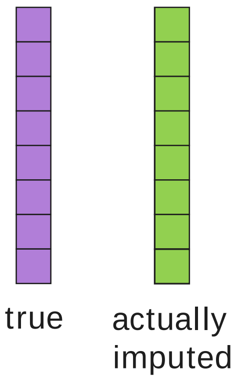
$$k_{j_2}^{(0)} = \max_{z\in\mathcal{T}_{j_2}} \left|F_{j_2}^{(true)}(z) - F_{j_2}^{(imp)}(z)\right|$$

Distributional accuracy measures (multivariate)

Kolmogorov-Smirnov-based approach

Now Cramér-Wold (multivariate): all (normed) mixtures of variables

KS for an artificial variable j_t



$$t = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0, \ldots, 0\right) \in \mathbb{R}^s$$

s: the number of metric and ordinal variables

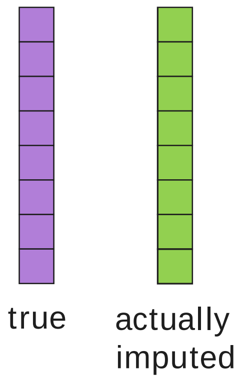
$$k_{j_t}^{(0)} = \max_{z \in \mathcal{T}_{j_t}} \left| F_{j_t}^{(true)}(z) - F_{j_t}^{(imp)}(z) \right|$$

Distributional accuracy measures (multivariate)

Kolmogorov-Smirnov-based approach

Now Cramér-Wold (multivariate): all (normed) mixtures of variables

KS for an artificial variable j_t



$$t = \left(\frac{1}{\sqrt{5}}, \frac{2}{\sqrt{5}}, 0, \dots, 0\right) \in \mathbb{R}^s$$

s: the number of metric and ordinal variables

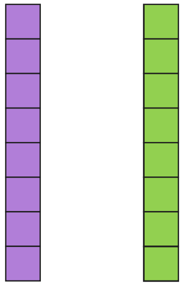
$$k_{j_t}^{(0)} = \max_{z \in \mathcal{T}_{j_t}} \left| F_{j_t}^{(true)}(z) - F_{j_t}^{(imp)}(z) \right|$$

Distributional accuracy measures (multivariate)

Kolmogorov-Smirnov-based approach

Now Cramér-Wold (multivariate): all (normed) mixtures of variables

KS for an artificial variable j_t



true actually imputed

$$t = \left(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, 0, \dots, 0\right) \in \mathbb{R}^s$$

s: the number of metric and ordinal variables

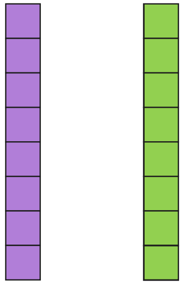
$$k_{j_t}^{(0)} = \max_{z \in \mathcal{T}_{j_t}} \left| F_{j_t}^{(true)}(z) - F_{j_t}^{(imp)}(z) \right|$$

Distributional accuracy measures (multivariate)

Kolmogorov-Smirnov-based approach

Now Cramér-Wold (multivariate): all (normed) mixtures of variables

KS for an artificial variable j_t



true

actually imputed

$$t = (1, 0, \ldots, 0) \in \mathbb{R}^s$$

s: the number of metric and ordinal variables

$$k_{j_t}^{(0)} = \max_{z \in \mathcal{T}_{j_t}} \left| F_{j_t}^{(true)}(z) - F_{j_t}^{(imp)}(z) \right|$$

Distributional accuracy measures (multivariate)

Kolmogorov-Smirnov-based approach

Now Cramér-Wold (multivariate): all (normed) mixtures of variables

- For all normed $t \in \mathbb{R}^s$ we now have

$$k_{j_t}^{(0)} = \max_{z \in \mathcal{T}_{j_t}} \left| F_{j_t}^{(true)}(z) - F_{j_t}^{(imp)}(z) \right|.$$

- They are then combined by the maximum or by the average over all those t to one (joint) test statistic. (Or multiple testing ...)
- The lower this joint test statistic is, the less evidence exists against the null hypothesis that the true and the original data set coincide.

Distributional accuracy measures (multivariate)

Kolmogorov-Smirnov-based approach

Now Cramér-Wold (multivariate): all (normed) mixtures of variables

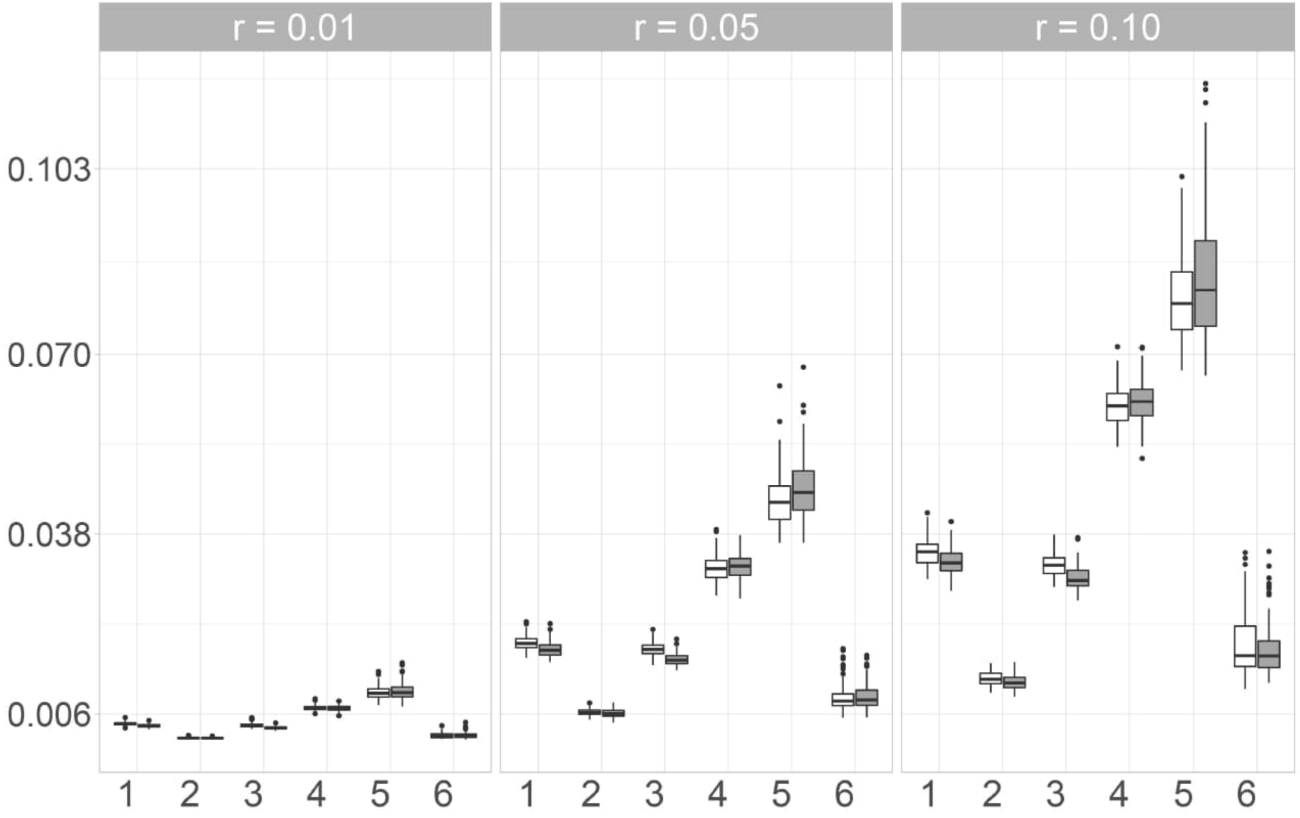
- A technical detail: In theory, for all normed $t \in \mathbb{R}^s$ we now have

$$k_{j_t}^{(0)} = \max_{z \in \mathcal{T}_{j_t}} \left| F_{j_t}^{(true)}(z) - F_{j_t}^{(imp)}(z) \right|.$$

- In practice: We can never go through all normed $t \in \mathbb{R}^s$ (there are uncountably many).
- → Stochastic approach, i.e. do the calculation for a lot of ts, e.g. 1000.

Distributional accuracy measures (multivariate)

Kolmogorov-Smirnov-based approach



$\max_t (KS_t)$

MC = 100

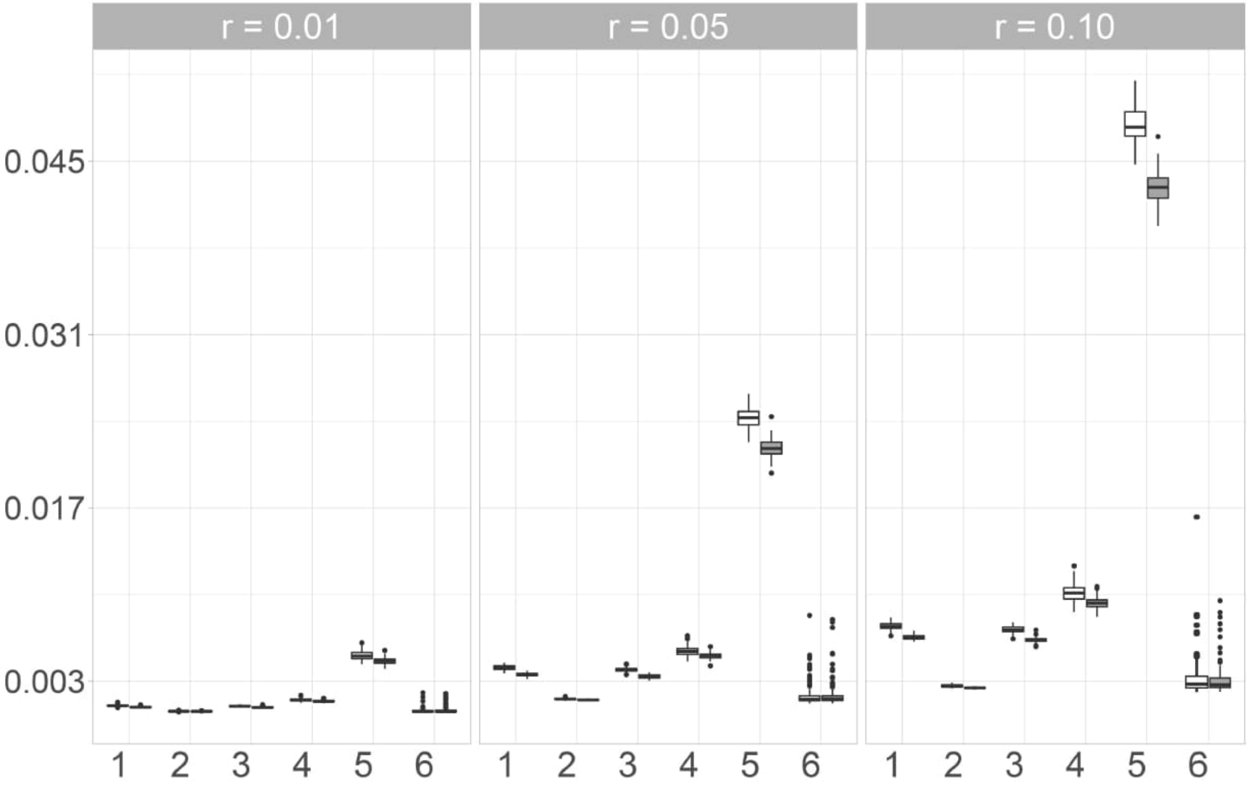
MCAR left,
MAR right

1: Amelia
2: Mice.Pmm
3: Mice.Norm
4: Mice.RF
5: Naive
6: missRanger

The lower, more or less, the better for distributional accuracy.

Distributional accuracy measures (multivariate)

Kolmogorov-Smirnov-based approach



$\underset{t}{\mathrm{ave}}\,(KS_t)$

MC = 100

MCAR left,
MAR right

1: Amelia
2: Mice.Pmm
3: Mice.Norm
4: Mice.RF
5: Naive
6: missRanger

The lower, more or less, the better for distributional accuracy.

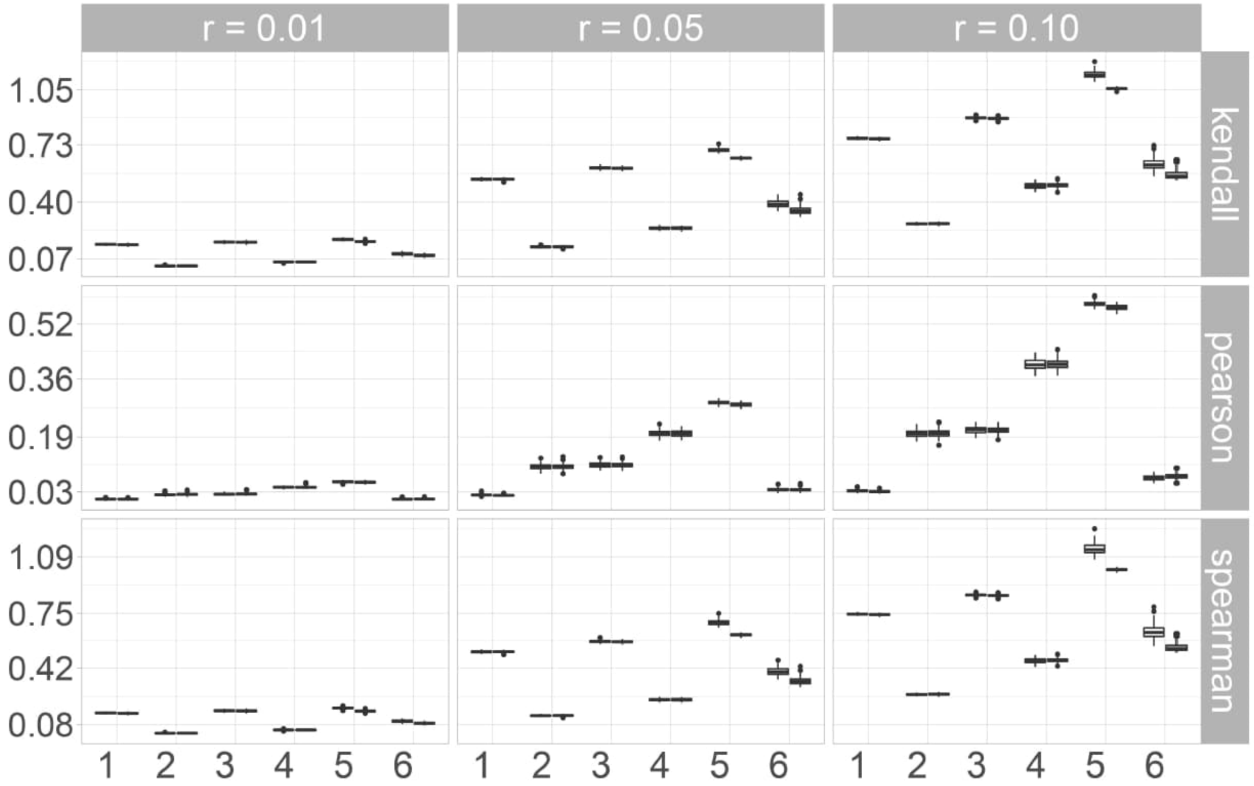
Distributional accuracy measures (multivariate)

Perhaps the simplest one: correlation-based approaches

- Compare the (different) correlation matrices of the variables
  - Pearson
  - Spearman $\rho_{Spearman}(X_{j_1}, X_{j_2}) = \rho_{Pearson}(F_{j_1}(X_{j_1}), F_{j_2}(X_{j_2}))$, empirically via ranks (distance)
  - Kendall theoretical relation to copulas, empirically via ranks (roughly)
- By Frobenius norm: $\left\| P^{true} - P^{imp} \right\|_F = \left(\sum_{i=1}^{s} \sum_{j=1}^{s} \left(\rho_{ij}^{true} - \rho_{ij}^{imp} \right)^2 \right)^{1/2}$
- By Maximum norm: $\left\| P^{true} - P^{imp} \right\|_{MAX} = \max_{i,j=1,\dots,s} \left| \rho_{ij}^{true} - \rho_{ij}^{imp} \right|$

Distributional accuracy measures (multivariate)

Perhaps the simplest one: correlation-based approaches



Frobenius norm

MC = 100

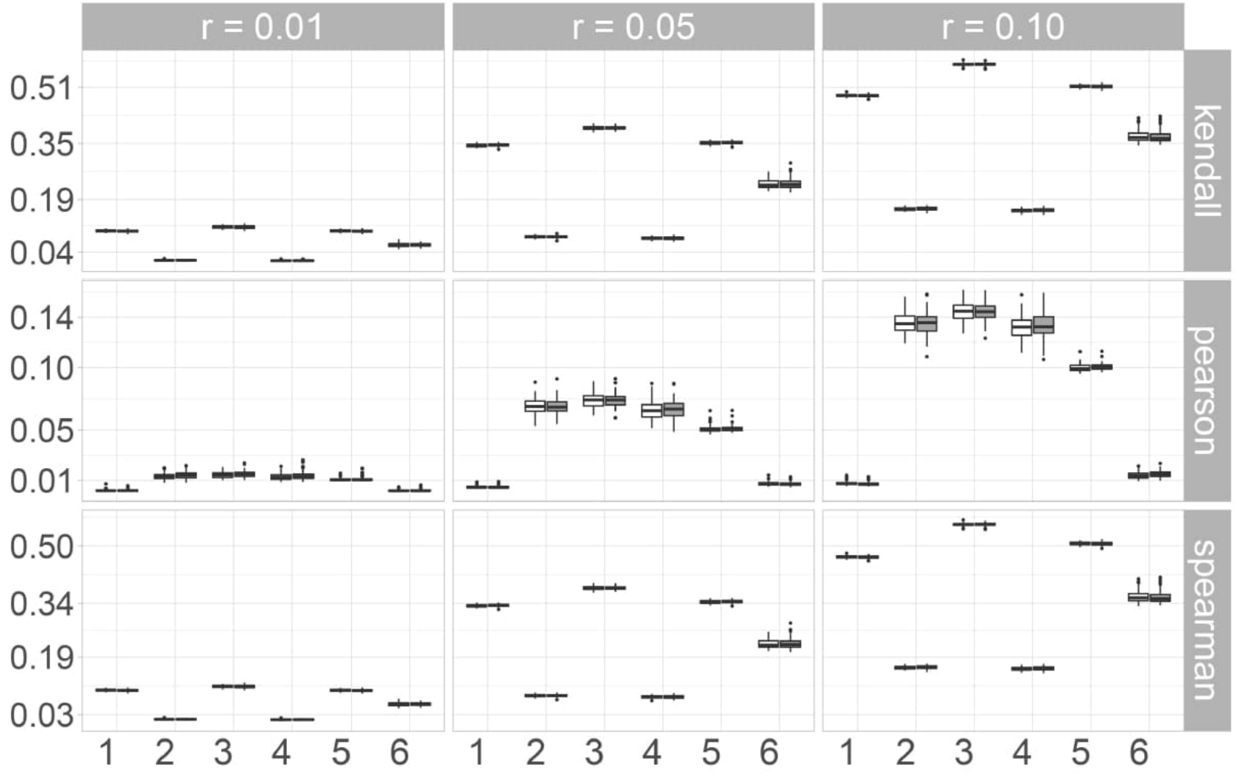
MCAR left,
MAR right

1: Amelia
2: Mice.Pmm
3: Mice.Norm
4: Mice.RF
5: Naive
6: missRanger

The lower, more or less, the better for distributional accuracy.

Distributional accuracy measures (multivariate)

Perhaps the simplest one: correlation-based approaches



Maximum norm

MC = 100

MCAR left,
MAR right

1: Amelia
2: Mice.Pmm
3: Mice.Norm
4: Mice.RF
5: Naive
6: missRanger

The lower, more or less, the better for distributional accuracy.

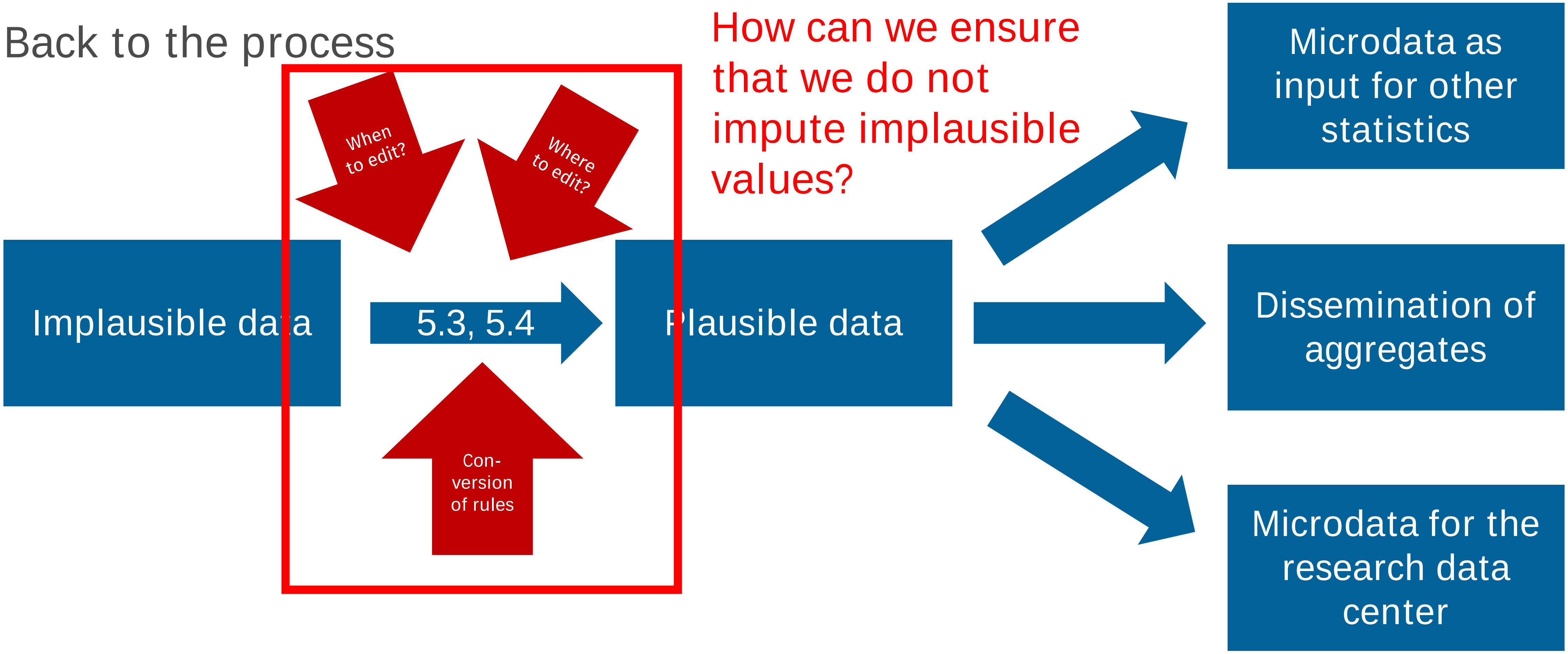
Accuracy measures

Open questions

- univariate:
  - distributional accuracy for nominal variables
- multivariate:
  - distributional accuracy for nominal variables
  - distributional accuracy for data sets with variables of different scales (levels of measurement)
  - curse of dimensionality
  - computational burden
- Standardisation / implementation: Further development of ImpACT? (Darren Gray, Marouane Seffal, Steffen Moritz)

Imputation under constraints

Back to the process



Imputation under constraints

A non-complete list of some ideas (better ones and worse ones)

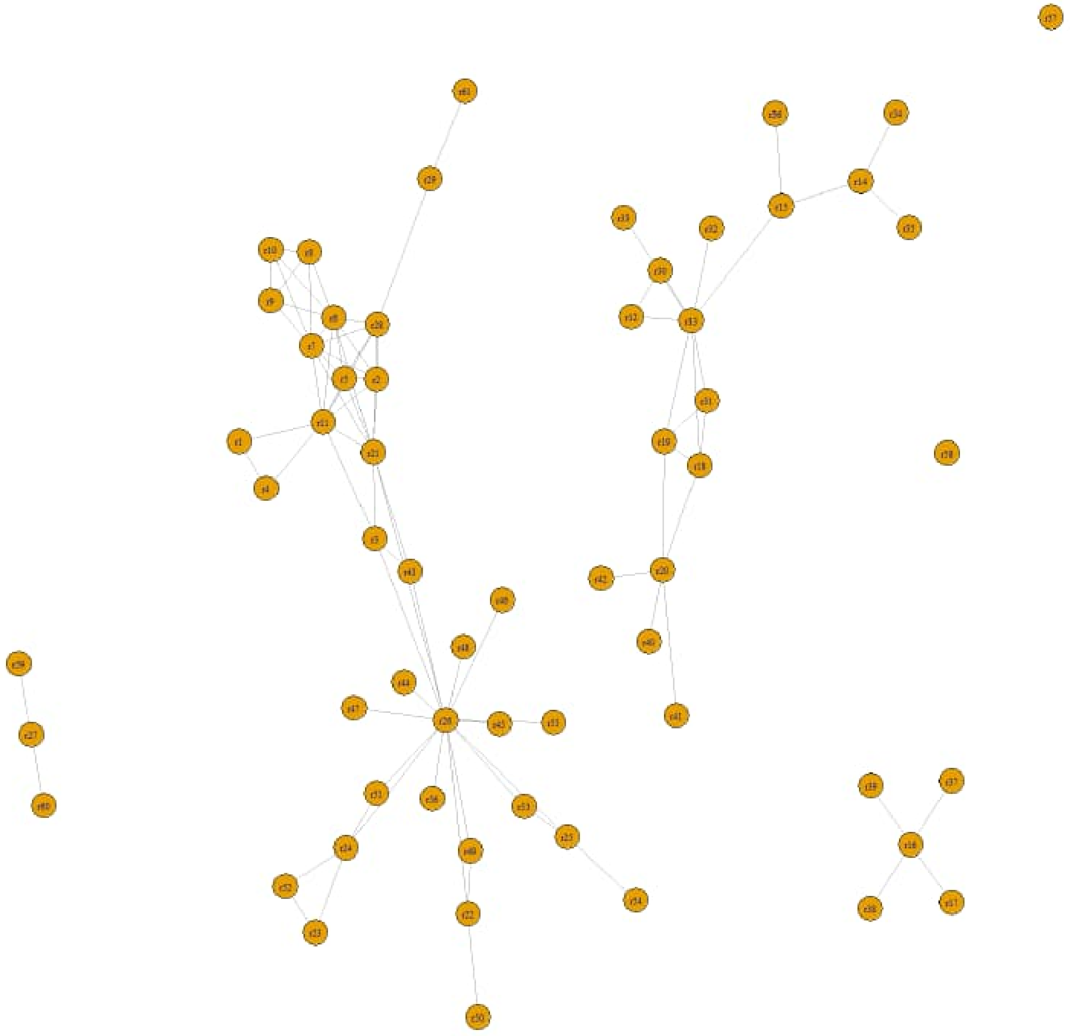
- Use a donor-based method (as e.g. CANCEIS is doing) and use only donors that fulfill all edit rules
- If values resulting from a regression-based method do not fulfill all edit rules, do not accept the value and draw again – until it works – if there is a theoretical chance that it works
- Model you regression-based method directly in a way that you produce only values that fulfill the edit rules

Imputation under constrair

Some insights into ongoing work

Empirical illustration

- Many variables (48), many rules (61), N = 17,286 observations
- Many rules are connected via variables
- Many variables are connected via rules



Aßmann C, Würbach A, Saidani Y, Dumpert F (2024) Full conditional distributions for handling restrictions in the context of automated statistical data editing. UNECE Expert Meeting on Statistical Data Editing, https://unece.org/sites/default/files/2024-09/SDE2024_S3_LIFBI_A%C3%9Fmann_D.pdf

Imputation under constraints

Some insights into ongoing work

- Focus not on all possible edit rules but on nested (equality and inequality) restrictions involving several variables
- Bayesian approach (→ nice also for estimating uncertainty)
- Including aspects like
  - censoring: a perceived continuous random variable has probability mass at one specific point that routinely would have a probability mass of zero
  - truncation: the range of the random variable is restricted to a range of values being element of an open interval

Aßmann C, Würbach A, Saidani Y, Dumpert F (2024) Full conditional distributions for handling restrictions in the context of automated statistical data editing. UNECE Expert Meeting on Statistical Data Editing, https://unece.org/sites/default/files/2024-09/SDE2024_S3_LIFBI_A%C3%9Fmann_D.pdf

Imputation under constraints

Some insights into ongoing work

- Key element: specification of full conditional distributions for the implausible values by univariate CART in order to take into account the dependencies among the variables (→ this also includes a lot of edit-rule-based dependencies, but not necessarily all)

- Decomposition of the joint density sequentially (not depending on the order):

$$f(X_1, \ldots, X_P|\theta) = f(X_1|\theta)f(X_2|X_1, \theta) \ldots f(X_P|X_1, \ldots, X_{P-1}, \theta)$$

- Decision: In most of the cases, we do not locate the error(s) among several variables involved but estimate all of them (under constraints)

Aßmann C, Würbach A, Saidani Y, Dumpert F (2024) Full conditional distributions for handling restrictions in the context of automated statistical data editing. UNECE Expert Meeting on Statistical Data Editing, https://unece.org/sites/default/files/2024-09/SDE2024_S3_LIFBI_A%C3%9Fmann_D.pdf

Imputation under constraints

Some insights into ongoing work

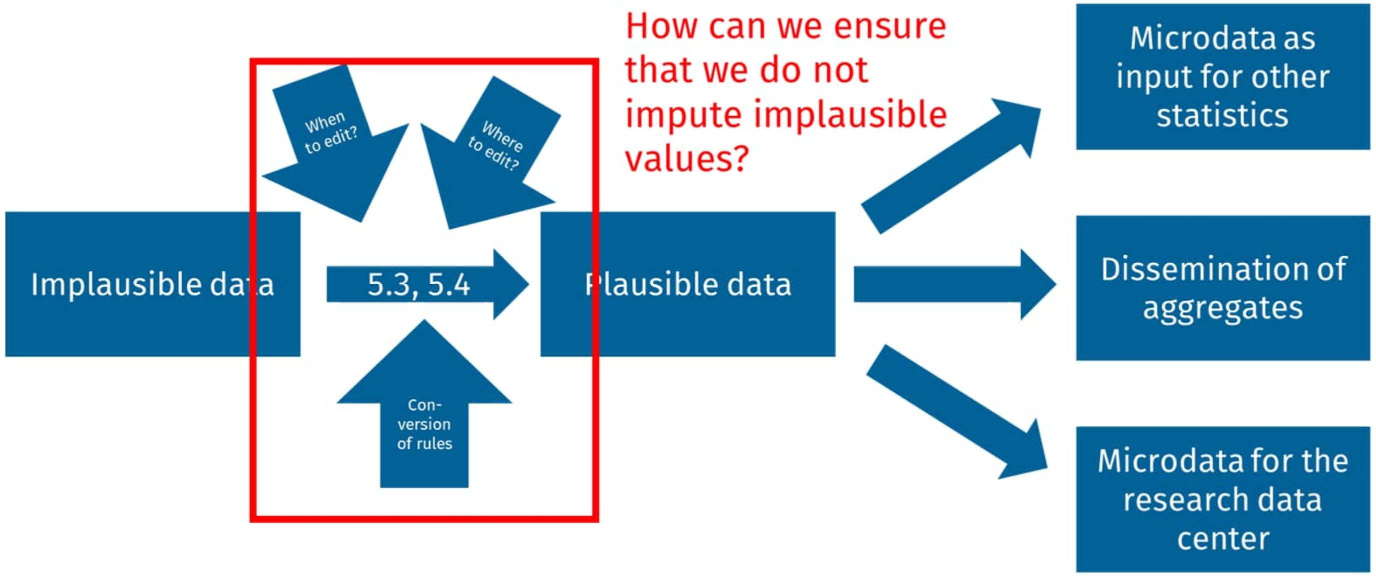
Simulation

- Observed observations deliver the initial conditional distributions, augmented observations the proceeding ones
- Missing values are then estimated per unit (not per variable), starting with the unit with the most sure (i.e. not missing) values
- CART was used to provide empirical distributions in its leaves with data D
  → Often in that example: Censored truncated normal distribution

$$f_{\mathcal{CTN}}(x|D) = p(D)I(x=0) + \left(1-p(D)\right)\frac{\phi\left(\frac{x-\mu(D)}{\sigma(D)}\right)}{\Phi\left(\frac{v^{(53)}-\mu(D)}{\sigma(D)}\right)-\Phi\left(\frac{-\mu(D)}{\sigma(D)}\right)}I(0<x<v^{(53)})$$

Imputation under constraints

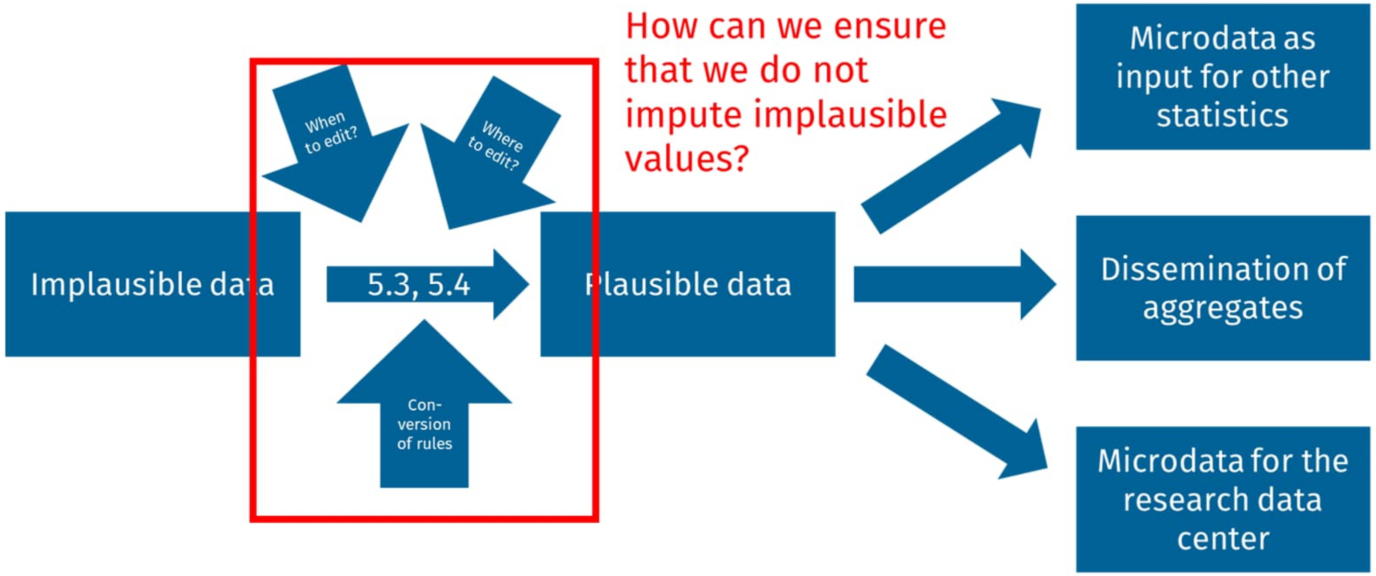
Conclusion



- The concept works in principle
- So far: a lot of manual work to decide to “autoEdit”
  - no automated translation of rules to conditional distributions
  - no implemented heuristic on the order of the imputation
- No error localization: We surely delete correct information

Imputation under constraints

To dos for the future



- Solve the problems from the slide before
- Combine the thoughts here with the evaluation measures
- Don’t give up ☺

Why do we do all this?

Quality of official statistics

- Aspects of the processes
  - Sound methodology, appropriate statistical procedures, non-excessive burden on respondents, cost effectiveness
- Aspects of the products
  - Relevance, accuracy and reliability, timeliness and punctuality, coherence and comparability, accessibility and clarity

https://ec.europa.eu/eurostat/web/quality/european-quality-standards/quality-assurance-framework

Destatis
Statistisches Bundesamt

Contact

Statistisches Bundesamt
65180 Wiesbaden
Germany

Florian Dumpert
florian.dumpert@destatis.de
Phone +49 611 75-3887

www.destatis.de

www.destatis.de/kontakt