Preprocessing

K. Gibert

Knowledge Engineering and Machine Learning group at

Intelligent Data Science and Artificial Intelligence Research Center

Universitat Politècnica de Catalunya, (IDEAI-UPC) Barcelona

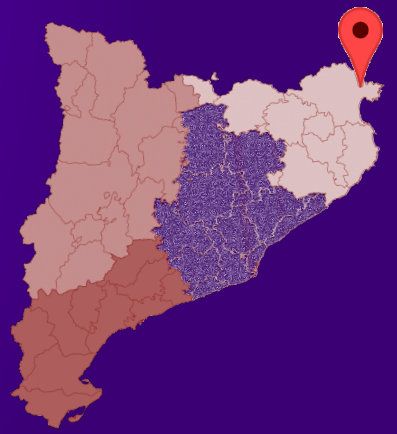
Dean of Illustrious Professional College of Informatics Engineering of Catalonia

President and founder of donesCOEINF

https://www.eio.upc.edu/en/homepages/karina,

karina.gibert@upc.edu, karina.gibert, @karinagibertk

Figuerenca



Abril, 2024

IDEAI UPC

First insight to Data

- Look at Metadata
- Determine rows and columns to be kept for the analysis
- Basic descriptive analysis of remanining variables
  - –Inspect anomalies, errors, missing data, outliers
- First report about data quality
- Preprocessing
- Verify after each processing step
- Final descriptive analysis *(report data improvements)*

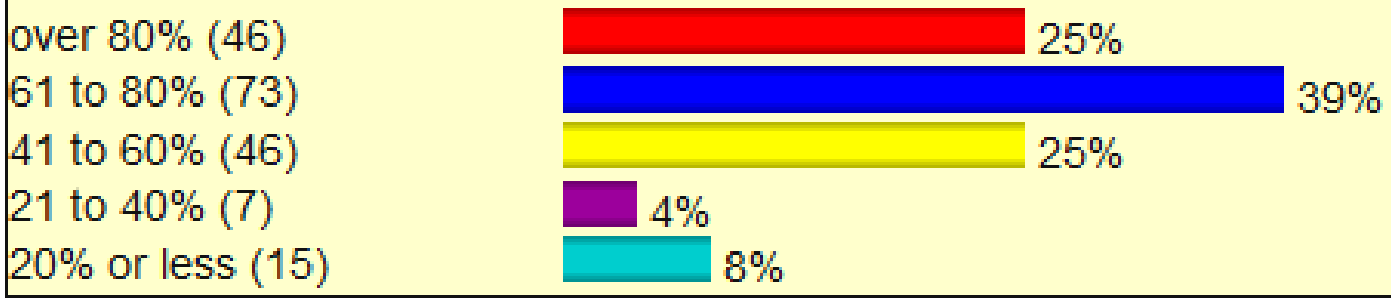
Impact of Preprocessing in real Data Mining projects

Data preparation part in data mining projects

Share Tweet

Poll

What % of time in your data mining project(s) is spent on data cleaning and preparation [187 votes total]



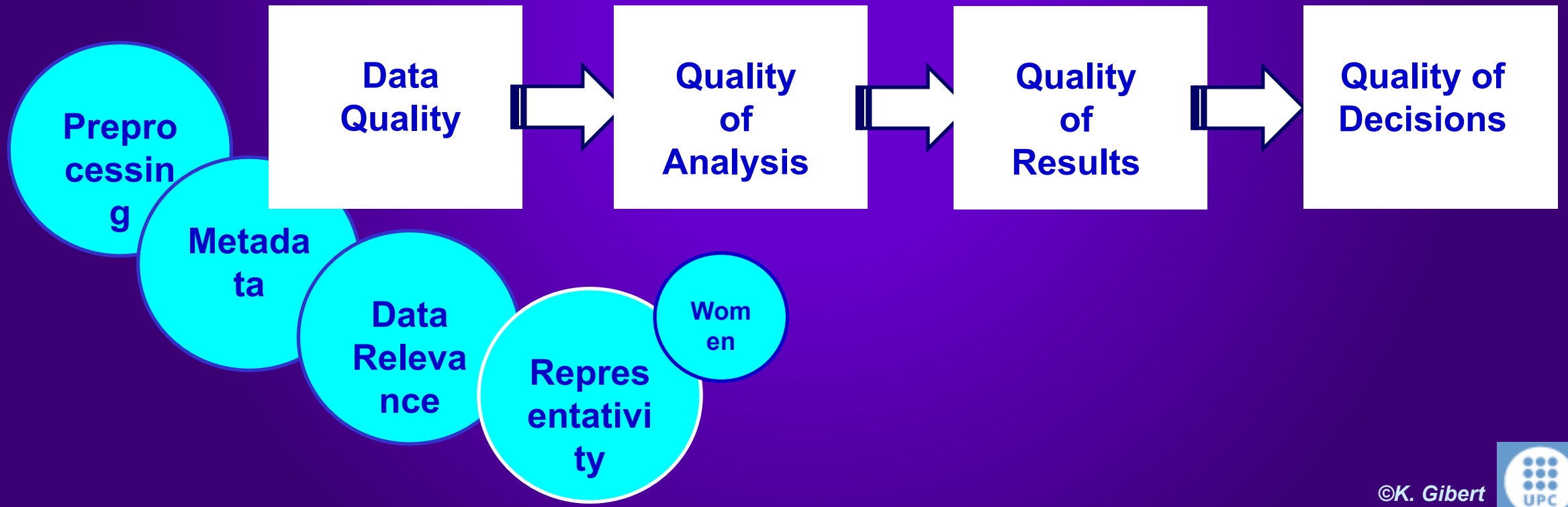
http://www.kdnuggets.com/polls/2003/data_preparation.htm

Data-driven decision/policy making

Data as a main source of information about reality

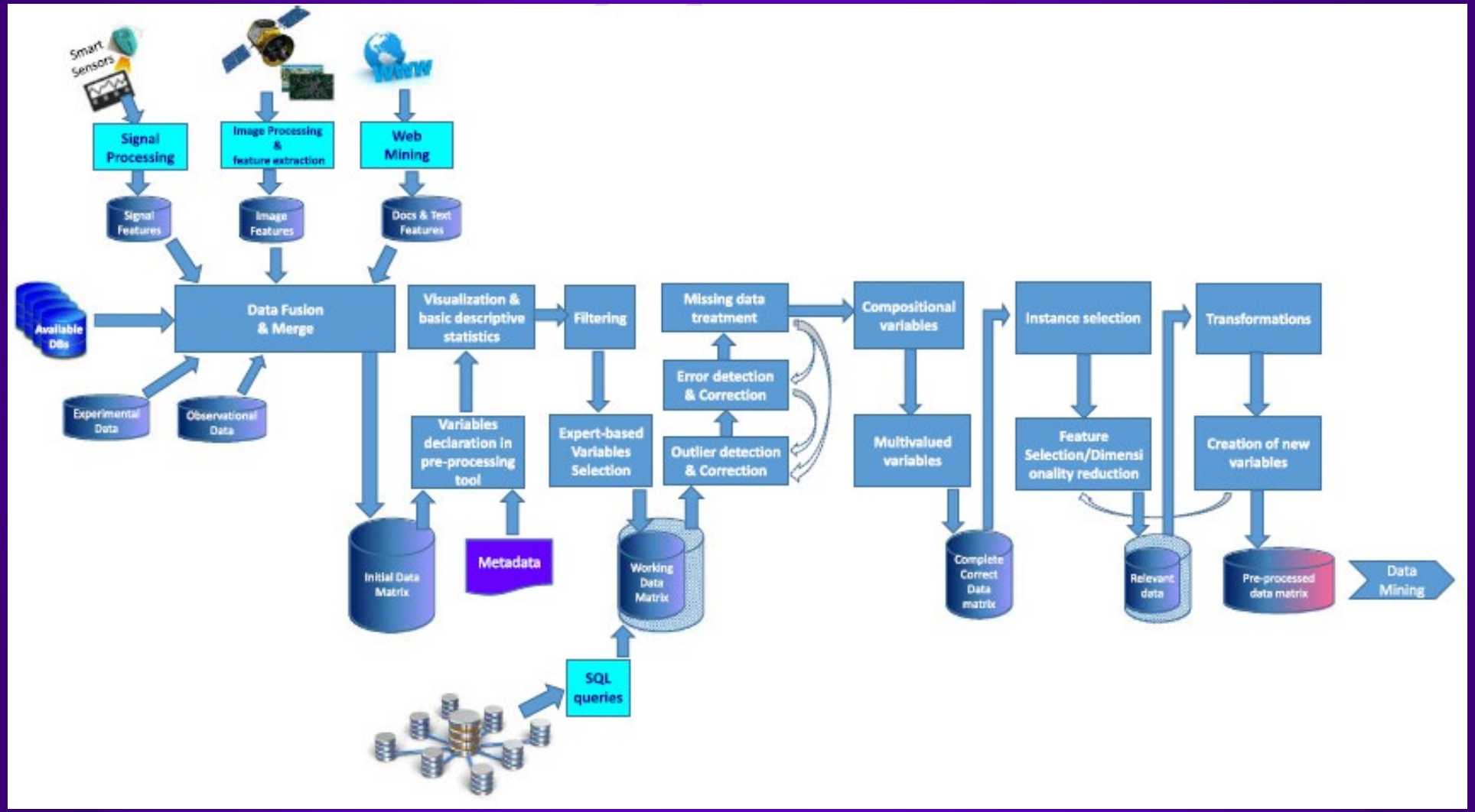
Core of Digital Transformation

Integration of data in decision-making proces



Methodology

[Gibert Aicomm2016]



Gibert, K., M. Sànchez-Marrè, J. Izquierdo (2016) A Survey on Pre-processing Techniques in the Context of Environmental Data Mining. Artificial Intelligence in Communications, 29(6): 627-663, IOSPress DOI: 10.3233/AIC-160710

Preprocessing

Data cleaning
Data preparation
Data preprocessing

- Formatting issues, building software context
- Determining working matrix, Filtering
- Identification and treatment of missing data
- Identification and treatment of outliers
- Identification and treatment of errors *(correct when possible)*
- Feature selection/extraction, dimensionality reduction
- Instance selection
- Data transformation
- Derivation of new variables

Determine Data Matrix

Which data matrix rows?

Define target population

Objects selection

Which data matrix columns?

Determine objects descritpion

Variables selection

Objects Selection

Inclusion/Exclusion criteria
Filtering

Select from a data base or data warehouse or from real individuals (costs are different)

- Experimental data (experimental design)
- Observational data (sample theory)

Define the target population

Determines scope of conclusions

Goals' oriented Variables Selection

- Often expert-guided
  (highly related with goal of analysis)

  - Be maximalists
    - Eliminate irrelevant or redundant information is less risky than detect lack of relevant things to be added in a second wave

  - Technically, to complete a final submatrix is highly costly (in both time and resources)

Reading Data
and
Variables Declaration

Reading and declaring data

- Verify that software got all rows and columns
  - Care with Spanish and English .csv files
- Verify that software understands variable types properly
  - Care with qualitative variables codified by integers
  - Care with numerical interpreted as textual variables
  - Metadata helps
- Ensure proper ordering in ordinal variables
- Use short modality labels

Reading and declaring data

Practical activity

R package preliminars

Integer
Double
Character
Logical (boolean)
Date
Vector and Matrix (numerical)
Factors (qualitative vectors)
Data Frames
Functions
Lists

Default creation with the assignment: <- or ->

Special values:

- NA: for missing data
- Inf: Infinity (division by 0…)
- NaN: error in the function results

Reading and declaring data

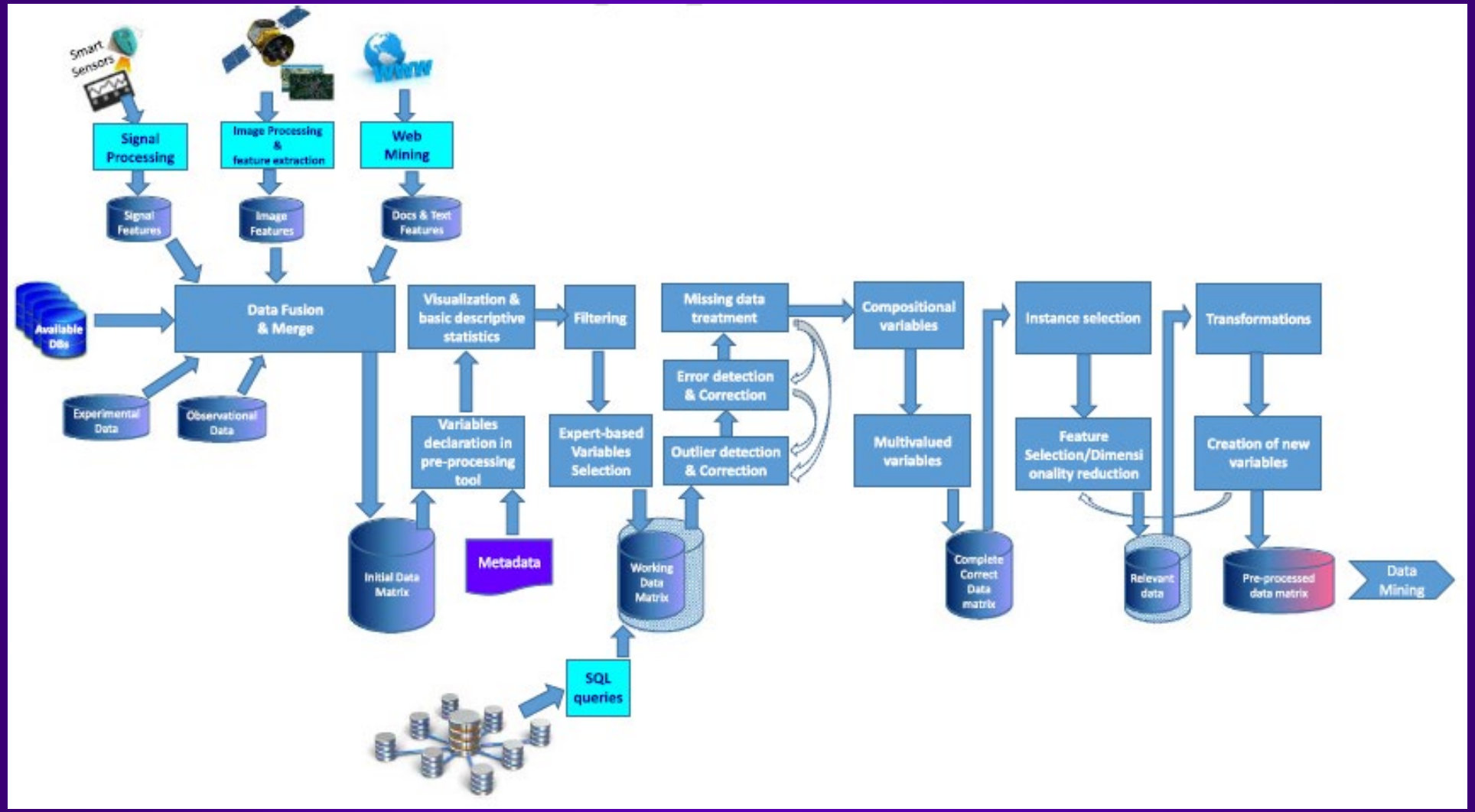
- Verify that software got all rows and columns
  - Care with Spanish and English .csv files
- Verify that software understands variable types properly
  - Care with qualitative variables codified by integers
  - Care with numerical interpreted as textual variables
  - Metadata helps
- Ensure proper ordering in ordinal variables
- Use short modality labels

Terminology normalization

- Accents
- Multilanguage
- *Hyphenation*
- *Likerts as qualy (with labels)*
- *When a variable is conditioned to another use ifs*

Methodology

[Gibert Aicomm2016]



Gibert, K., M. Sànchez-Marrè, J. Izquierdo (2016) A Survey on Pre-processing Techniques in the Context of Environmental Data Mining. Artificial Intelligence in Communications, 29(6): 627-663, IOSPress DOI: 10.3233/AIC-160710

Preprocessing

Data cleaning
Data preparation
Data preprocessing

- Formatting issues, building software context
- Determining working matrix, Filtering
- Identification and treatment of missing data
- Identification and treatment of outliers
- Identification and treatment of errors *(correct when possible)*
- Feature selection/extraction, dimensionality reduction
- Instance selection
- Data transformation
- Derivation of new variables

Missing data

(empty cells in data matrix)

- Types and diagnosis
- Little's test
- A simple descriptive alternative
- Some methods
  - Knn
  - The MIMMI method
  - MICE
  - Interpolation (for time series)

Missing data

(empty cells in data matrix)

- Randon missing
  - non problematic
  - casual
  - follow same distribution as present data
  - inputation is easy: mean, 0
- Non random missing: absence is informative
  - come from some particular part of population
  - probably correspond to special values
  - difficult to induce from the present data
  - inputation is much difficult
  - very critical
  - very dangerous to ignore those individuals
  - asking religion in israel (muslims do not answer)
  - Asking age to a lady over 45
  - Frequency of observations (microbio tests in water)
- Non applicable value (non-random, structural)
  - salary of a non-working person
  - number of pregnancies of a man
  - number of cigarretes of a non-smoker person
  - age of menopause

Missing data

Diagnoses

Little's MCAR test

H_0: Missings are completely at random (MCAR)
H_1: Missings are not random

$$d^2 = \sum_{j=1}^{J} n_j (\overline{X}_j - \overline{X}_j^*)^T \frac{1}{\hat{\sum}_j} (\overline{X}_j - \overline{X}_j^*) \sim \chi^2_{\sum r_j - K}$$

j=1:J missing patterns (subsets of missing variables in a case)
n_j *cases in missing pattern j*
$\overline{X}_j$ *maximum likelihood estimates of the grand means*
$\overline{X}_j^*$ *means local to cases in missing pattern*
$\hat{\sum}$ *maximum likelihood estimate of the covariance matrix*
rj number of complete variables for pattern j
K total number of variables

Searches significant differences in means conditioned to a certain subset of missing variables (pattern j)

Dades UNICEF 2004 http://www.unicef.org/spanish/infobycountry/index.html missing patterns

1	Països	Zona	PobTotal	INBzcap	EsperVida	%NounatsBaixPes	UsInstalSaneja	%VIH	%coneixPresHomes	$ConeixPresDones	AlfabetHomes	AlfabetDones	Mortalitat	Natalitat	Economia	despesaSalut/Defensa	Règim
2	CAMBOYA	AsiaOr+Pacific	13798	320	57	11	16	170		64	85	64	11	31	enDesenvolupament		Monarquia
3	COREA	AsiaOr+Pacific	22384		63	7	59						11	16	enDesenvolupament		Republica
4	FIJI	AsiaOr+Pacific	841	2690	68	10	98	0,6			94	91	6	23	enDesenvolupament	MMesDefensa	Republica
5	FILIPINES	AsiaOr+Pacific	81617	1170	71	20	73	9	59	44	93	93	5	25	enDesenvolupament	MesSalut	Republica
6	ILLES COOK	AsiaOr+Pacific	18			3	100								Subdesenvolupat		Monarquia
7	ILLES MARSHALL	AsiaOr+Pacific	60	2370		12	82								Subdesenvolupat		Republica
8	ILLES SALOMÓ	AsiaOr+Pacific	466	550	63	13	31						7	33	Subdesenvolupat		Monarquia
9	INDONÈSIA	AsiaOr+Pacific	220077	1140	67	9	52	110		23	92	83	7	21	enDesenvolupament	MesSalut	Republica
10	KIRIBATI	AsiaOr+Pacific	97	970		5	39								Subdesenvolupat		Republica
11	LAO, REPÚBLICA DEMOCRÀTICA	AsiaOr+Pacific	5792	390	55	14	24	1,7			77	61	12	35	enDesenvolupament		Republica
12	MALÀSIA	AsiaOr+Pacific	24894	4650	73	9		52			92	85	5	22	enDesenvolupament	MesSalut	Monarquia
13	MICRONÈSIA	AsiaOr+Pacific	110	1990	68	18	28						6	31	Subdesenvolupat		Republica
14	MONGÒLIA	AsiaOr+Pacific	2614	590	65	7	59	0,5		77	98	98	7	22	enDesenvolupament	MesSalut	Republica
15	MYANMAR	AsiaOr+Pacific	50004	220	61	15	73	330			94	86	10	20	enDesenvolupament	MesSalut	Dictadura
16	NAURU	AsiaOr+Pacific	13												Subdesenvolupat		Republica
17	NIUE	AsiaOr+Pacific	1			0	100								Subdesenvolupat		Monarquia
18	PALAU	AsiaOr+Pacific	20	6870		9	83								Subdesenvolupat		Republica
19	PAPUA NOVA GUINEA	AsiaOr+Pacific	5772	580	56	11	45	16			63	51	10	30	enDesenvolupament	MMesDefensa	Monarquia
20	SAMOA	AsiaOr+Pacific	184	1860	71	4	100				99	98	6	28	Subdesenvolupat		Monarquia
21	SINGAPUR	AsiaOr+Pacific	4273	24220	79	8		4,1			97	89	5	9	enDesenvolupament	MesSalut	Republica
22	TAILÀNDIA	AsiaOr+Pacific	63694	2540	70	9	99	0,2			95	91	7	16	enDesenvolupament	MMesDefensa	Monarquia
23	TIMOR-LESTE	AsiaOr+Pacific	887	550	56	12	33			6			12	50	Subdesenvolupat		Republica
24	TOKELAU	AsiaOr+Pacific													Subdesenvolupat		Monarquia
25	TONGA	AsiaOr+Pacific	102	1830	72	0	97				99	99	6	24	Subdesenvolupat		Monarquia
26	TUVALU	AsiaOr+Pacific	10			5	88								Subdesenvolupat		Monarquia
27	VANUATU	AsiaOr+Pacific	207	1340	69	6	50						6	31	enDesenvolupament		Republica
28	VIETNAM	AsiaOr+Pacific	83123	550	71	9	41	220		60	94	87	6	20	enDesenvolupament		Republica
29	XINA	AsiaOr+Pacific	1307989	1290	72	4	44	840			95	87	7	13	enDesenvolupament	MesSalut	Republica
30	ANGOLA	AfricaOr+Merid	15490	1030	41	12	30	240			82	54	22	48	enDesenvolupament	MesSalut	Republica
31	BOTSWANA	AfricaOr+Merid	1769	4340	35	10	41	350	90	93	76	82	27	26	enDesenvolupament	MesSalut	Republica
32	BURUNDI	AfricaOr+Merid	7282	90	44	16	36	250		47	67	52	19	45	enDesenvolupament	MesSalut	Republica
33	COMORAS	AfricaOr+Merid	777	530	64	25	23			41	63	49	7	36	enDesenvolupament		Republica
34	ERITREA	AfricaOr+Merid	4232	180	54	21	9	60		62			11	39	enDesenvolupament		Republica
35	ETIOPIA	AfricaOr+Merid	75600	110	48	15	6	1500			49	34	16	41	enDesenvolupament	MesSalut	Republica
36	KENYA	AfricaOr+Merid	33467	460	48	10	48	1200	68	59	78	70	15	39	enDesenvolupament	MMesDefensa	Republica
37	LESOTHO	AfricaOr+Merid	1798	740	35	14	37	320		58	74	90	25	28	enDesenvolupament	MMesDefensa	Monarquia
38	MADAGASCAR	AfricaOr+Merid	18113	300	56	17	33	140	54	49	76	65	12	39	enDesenvolupament	MMesDefensa	Republica

bert

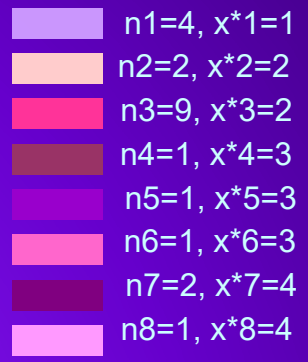
UPC

IDEAI

Dades UNICEF 2004 http://www.unicef.org/spanish/infobycountry/index.html

Missing pattern 1: (ConeixPresHomes)
Missing pattern 2: (ConeixPresHomes+DespesaSalut/Defensa)
Missing pattern 3: (ConeixPresHomes+ConeixPresDones)
Missing pattern 4: (ConeixPresHomes+ConeixPresDones+DespSal/Def)
Missing pattern 5: (ConeixPresHomes+AlfabetsDones+DespSal/Def)
Missing pattern 6: (VIH+ConeixPresHomes+ DespSal/Def)
Missing pattern 7: (VIH+ConeixPresHomes+ ConeixPresHomes+ConeixPresDones+DespSal/Def)
Missing pattern 8: (ConeixPresHomes+AlfabetHomes+AlfabetDones+DespSal/Def)

missing patterns



	Països	Zona	PobTotal	INBzcap	EsperVida	%NounatsBaixPes	UsInstalSaneja	VIH	%coneixPresHomes	$ConeixPresDones	AlfabetHomes	AlfabetDones	Mortalitat	Natalitat	Economia	despesaSalut/Defensa	Règim
2	CAMBOYA	AsiaOr+Pacific	13798	320	57	11	16	170		64	85	64	11	31	enDesenvolupament		Monarquia
3	COREA	AsiaOr+Pacific	22384		63	7	59						11	16	enDesenvolupament		Republica
4	FIJI	AsiaOr+Pacific	841	2690	68	10	98	0,6			94	91	6	23	enDesenvolupament	MMesDefensa	Republica
5	FILIPINES	AsiaOr+Pacific	81617	1170	71	20	73	9	59	44	93	93	5	25	enDesenvolupament	MesSalut	Republica
6	ILLES COOK	AsiaOr+Pacific	18			3	100								Subdesenvolupat		Monarquia
7	ILLES MARSHALL	AsiaOr+Pacific	60	2370		12	82								Subdesenvolupat		Republica
8	ILLES SALOMÓ	AsiaOr+Pacific	466	550	63	13	31						7	33	Subdesenvolupat		Monarquia
9	INDONÈSIA	AsiaOr+Pacific	220077	1140	67	9	52	110		23	92	83	7	21	enDesenvolupament	MesSalut	Republica
10	KIRIBATI	AsiaOr+Pacific	97	970		5	39								Subdesenvolupat		Republica
11	LAO, REPÚBLICA DEMOCRÀTICA	AsiaOr+Pacific	5792	390	55	14	24	1,7			77	61	12	35	enDesenvolupament		Republica
12	MALÀSIA	AsiaOr+Pacific	24894	4650	73	9		52			92	85	5	22	enDesenvolupament	MesSalut	Monarquia
13	MICRONÈSIA	AsiaOr+Pacific	110	1990	68	18	28						6	31	Subdesenvolupat		Republica
14	MONGÒLIA	AsiaOr+Pacific	2614	590	65	7	59	0,5		77	98	98	7	22	enDesenvolupament	MesSalut	Republica
15	MYANMAR	AsiaOr+Pacific	50004	220	61	15	73	330			94	86	10	20	enDesenvolupament	MesSalut	Dictadura
16	NAURU	AsiaOr+Pacific	13												Subdesenvolupat		Republica
17	NIUE	AsiaOr+Pacific	1			0	100								Subdesenvolupat		Monarquia
18	PALAU	AsiaOr+Pacific	20	6870		9	83								Subdesenvolupat		Republica
19	PAPUA NOVA GUINEA	AsiaOr+Pacific	5772	580	56	11	45	16			63	51	10	30	enDesenvolupament	MMesDefensa	Monarquia
20	SAMOA	AsiaOr+Pacific	184	1860	71	4	100				99	98	6	28	Subdesenvolupat		Monarquia
21	SINGAPUR	AsiaOr+Pacific	4273	24220	79	8		4,1			97	89	5	9	enDesenvolupament	MesSalut	Republica
22	TAILÀNDIA	AsiaOr+Pacific	63694	2540	70	9	99	0,2			95	91	7	16	enDesenvolupament	MMesDefensa	Monarquia
23	TIMOR-LESTE	AsiaOr+Pacific	887	550	56	12	33			6			12	50	Subdesenvolupat		Republica
24	TOKELAU	AsiaOr+Pacific													Subdesenvolupat		Monarquia
25	TONGA	AsiaOr+Pacific	102	1830	72	0	97				99	99	6	24	Subdesenvolupat		Monarquia
26	TUVALU	AsiaOr+Pacific	10			5	88								Subdesenvolupat		Monarquia
27	VANUATU	AsiaOr+Pacific	207	1340	69	6	50						6	31	enDesenvolupament		Republica
28	VIETNAM	AsiaOr+Pacific	83123	550	71	9	41	220		60	94	87	6	20	enDesenvolupament		Republica
29	XINA	AsiaOr+Pacific	1307989	1290	72	4	44	840			95	87	7	13	enDesenvolupament	MesSalut	Republica
30	ANGOLA	AfricaOr+Merid	15490	1030	41	12	30	240			82	54	22	48	enDesenvolupament	MesSalut	Republica
31	BOTSWANA	AfricaOr+Merid	1769	4340	35	10	41	350	90	93	76	82	27	26	enDesenvolupament	MesSalut	Republica
32	BURUNDI	AfricaOr+Merid	7282	90	44	16	36	250		47	67	52	19	45	enDesenvolupament	MesSalut	Republica
33	COMORAS	AfricaOr+Merid	777	530	64	25	23			41	63	49	7	36	enDesenvolupament		Republica
34	ERITREA	AfricaOr+Merid	4232	180	54	21	9	60		62			11	39	enDesenvolupament		Republica
35	ETIOPIA	AfricaOr+Merid	75600	110	48	15	6	1500			49	34	16	41	enDesenvolupament	MesSalut	Republica
36	KENYA	AfricaOr+Merid	33467	460	48	10	48	1200	68	59	78	70	15	39	enDesenvolupament	MMesDefensa	Republica
37	LESOTHO	AfricaOr+Merid	1798	740	35	14	37	320		58	74	90	25	28	enDesenvolupament	MMesDefensa	Monarquia
38	MADAGASCAR	AfricaOr+Merid	18113	300	56	17	33	140	54	49	76	65	12	39	enDesenvolupament	MMesDefensa	Republica

Dades UNICEF 2004 http://www.unicef.org/spanish/infobycountry/index.html

n1=4, x*1=1
n2=2, x*2=2
n3=9, x*3=2
n4=1, x*4=3
n5=1, x*5=3
n6=1, x*6=3
n7=2, x*7=4
n8=1, x*8=4

Missing pattern 9: (VIH+CPH+ CPD+AlfaH+AlfaD+DespSal/Def)
Missing pattern 10: (INB+VIH+CPH+ CPD+AlfaH+AlfaD+DespSal/Def))
Missing pattern 1: (EV+ VIH+CPH+ CPD+AlfaH+AlfaD+M+N+D/Def)
Missing pattern 12: (INB+ EV+ VIH+CPH+ CPD+AlfaH+AlfaD+M+N+D/Def)
Missing pattern 13: (INB+ EV+NBp+IS+ VIH+CPH+ CPD+AlfaH+AlfaD+M+N+D/Def)
Missing pattern 14: (Pob+INB+ EV+NBp+IS+ VIH+CPH+ CPD+AlfaH+AlfaD+M+N+D/Def)

missing patterns

n9=3, x*9=6
n10=1, x*10=7
n11=3, x*11=9
n12=3, x*12=10
n13=1, x*13=12
n14=1, x*14=13

1	Països	Zona	PobTotal	INBzcap	Esper Vida	%NounatsBaixPes	UsInstal Saneja	VIH	%coneixPresHomes	$ConeixPresDones	Alfabet Homes	Alfabet Dones	Mortalitat	Natalitat	Economia	despesaSalut/Defensa	Règim
2	CAMBOYA	AsiaOr+Pacific	13798	320	57	11	16	170		64	85	64	11	31	enDesenvolupament		Monarquia
3	COREA	AsiaOr+Pacific	22384		63	7	59						11	16	enDesenvolupament		Republica
4	FIJI	AsiaOr+Pacific	841	2690	68	10	98	0,6			94	91	6	23	enDesenvolupament	MMesDefensa	Republica
5	FILIPINES	AsiaOr+Pacific	81617	1170	71	20	73	9	59	44	93	93	5	25	enDesenvolupament	MesSalut	Republica
6	ILLES COOK	AsiaOr+Pacific	18			3	100								Subdesenvolupat		Monarquia
7	ILLES MARSHALL	AsiaOr+Pacific	60	2370		12	82								Subdesenvolupat		Republica
8	ILLES SALOMÓ	AsiaOr+Pacific	466	550	63	13	31						7	33	Subdesenvolupat		Monarquia
9	INDONÈSIA	AsiaOr+Pacific	220077	1140	67	9	52	110		23	92	83	7	21	enDesenvolupament	MesSalut	Republica
10	KIRIBATI	AsiaOr+Pacific	97	970		5	39								Subdesenvolupat		Republica
11	LAO, REPÚBLICA DEMOCRÀTICA	AsiaOr+Pacific	5792	390	55	14	24	1,7			77	61	12	35	enDesenvolupament		Republica
12	MALÀSIA	AsiaOr+Pacific	24894	4650	73	9		52			92	85	5	22	enDesenvolupament	MesSalut	Monarquia
13	MICRONÈSIA	AsiaOr+Pacific	110	1990	68	18	28						6	31	Subdesenvolupat		Republica
14	MONGÒLIA	AsiaOr+Pacific	2614	590	65	7	59	0,5		77	98	98	7	22	enDesenvolupament	MesSalut	Republica
15	MYANMAR	AsiaOr+Pacific	50004	220	61	15	73	330			94	86	10	20	enDesenvolupament	MesSalut	Dictadura
16	NAURU	AsiaOr+Pacific	13												Subdesenvolupat		Republica
17	NIUE	AsiaOr+Pacific	1			0	100								Subdesenvolupat		Monarquia
18	PALAU	AsiaOr+Pacific	20	6870		9	83								Subdesenvolupat		Republica
19	PAPUA NOVA GUINEA	AsiaOr+Pacific	5772	580	56	11	45	16			63	51	10	30	enDesenvolupament	MMesDefensa	Monarquia
20	SAMOA	AsiaOr+Pacific	184	1860	71	4	100				99	98	6	28	Subdesenvolupat		Monarquia
21	SINGAPUR	AsiaOr+Pacific	4273	24220	79	8		4,1			97	89	5	9	enDesenvolupament	MesSalut	Republica
22	TAILÀNDIA	AsiaOr+Pacific	63694	2540	70	9	99	0,2			95	91	7	16	enDesenvolupament	MMesDefensa	Monarquia
23	TIMOR-LESTE	AsiaOr+Pacific	887	550	56	12	33			6			12	50	Subdesenvolupat		Republica
24	TOKELAU	AsiaOr+Pacific													Subdesenvolupat		Monarquia
25	TONGA	AsiaOr+Pacific	102	1830	72	0	97				99	99	6	24	Subdesenvolupat		Monarquia
26	TUVALU	AsiaOr+Pacific	10			5	88								Subdesenvolupat		Monarquia
27	VANUATU	AsiaOr+Pacific	207	1340	69	6	50						6	31	enDesenvolupament		Republica
28	VIETNAM	AsiaOr+Pacific	83123	550	71	9	41	220		60	94	87	6	20	enDesenvolupament		Republica
29	XINA	AsiaOr+Pacific	1307989	1290	72	4	44	840			95	87	7	13	enDesenvolupament	MesSalut	Republica
30	ANGOLA	AfricaOr+Merid	15490	1030	41	12	30	240			82	54	22	48	enDesenvolupament	MesSalut	Republica
31	BOTSWANA	AfricaOr+Merid	1769	4340	35	10	41	350	90	93	76	82	27	26	enDesenvolupament	MesSalut	Republica
32	BURUNDI	AfricaOr+Merid	7282	90	44	16	36	250		47	67	52	19	45	enDesenvolupament	MesSalut	Republica
33	COMORAS	AfricaOr+Merid	777	530	64	25	23			41	63	49	7	36	enDesenvolupament		Republica
34	ERITREA	AfricaOr+Merid	4232	180	54	21	9	60		62			11	39	enDesenvolupament		Republica
35	ETIOPIA	AfricaOr+Merid	75600	110	48	15	6	1500			49	34	16	41	enDesenvolupament	MesSalut	Republica
36	KENYA	AfricaOr+Merid	33467	460	48	10	48	1200	68	59	78	70	15	39	enDesenvolupament	MMesDefensa	Republica
37	LESOTHO	AfricaOr+Merid	1798	740	35	14	37	320		58	74	90	25	28	enDesenvolupament	MMesDefensa	Monarquia
38	MADAGASCAR	AfricaOr+Merid	18113	300	56	17	33	140	54	49	76	65	12	39	enDesenvolupament	MMesDefensa	Republica

Dades UNICEF 2004 http://www.unicef.org/spanish/infobycountry/index.html

_missing patterns

- n1=4, x*1=1
- n2=2, x*2=2
- n3=9, x*3=2
- n4=1, x*4=3
- n5=1, x*5=3
- n6=1, x*6=3
- n7=2, x*7=4
- n8=1, x*8=4
- n9=3, x*9=6
- n10=1, x*10=7
- n11=3, x*11=9
- n12=3, x*12=10
- n13=1, x*13=12
- n14=1, x*14=13

15 patterns

	Països	Zona	PobTotal	INBzcap	Esper Vida	%NounatsBaixPes	UsInstal Saneja	%VIH	%coneixPresHomes	$ConeixPresDones	Alfabet Homes	Alfabet Dones	Mortalitat	Natalitat	Economia	despesaSalut/Defensa	Règim
2	CAMBOYA	AsiaOr+Pacific	13798	320	57	11	16	170		64	85	64	11	31	enDesenvolupament		Monarquia
3	COREA	AsiaOr+Pacific	22384		63	7	59						11	16	enDesenvolupament		Republica
4	FIJI	AsiaOr+Pacific	841	2690	68	10	98	0,6			94	91	6	23	enDesenvolupament	MMesDefensa	Republica
5	FILIPINES	AsiaOr+Pacific	81617	1170	71	20	73	9	59	44	93	93	5	25	enDesenvolupament	MesSalut	Republica
6	ILLES COOK	AsiaOr+Pacific	18			3	100								Subdesenvolupat		Monarquia
7	ILLES MARSHALL	AsiaOr+Pacific	60	2370		12	82								Subdesenvolupat		Republica
8	ILLES SALOMÓ	AsiaOr+Pacific	466	550	63	13	31						7	33	Subdesenvolupat		Monarquia
9	INDONÈSIA	AsiaOr+Pacific	220077	1140	67	9	52	110		23	92	83	7	21	enDesenvolupament	MesSalut	Republica
10	KIRIBATI	AsiaOr+Pacific	97	970		5	39								Subdesenvolupat		Republica
11	LAO, REPÚBLICA DEMOCRÀTICA	AsiaOr+Pacific	5792	390	55	14	24	1,7			77	61	12	35	enDesenvolupament		Republica
12	MALÀSIA	AsiaOr+Pacific	24894	4650	73	9		52			92	85	5	22	enDesenvolupament	MesSalut	Monarquia
13	MICRONÈSIA	AsiaOr+Pacific	110	1990	68	18	28						6	31	Subdesenvolupat		Republica
14	MONGÒLIA	AsiaOr+Pacific	2614	590	65	7	59	0,5		77	98	98	7	22	enDesenvolupament	MesSalut	Republica
15	MYANMAR	AsiaOr+Pacific	50004	220	61	15	73	330			94	86	10	20	enDesenvolupament	MesSalut	Dictadura
16	NAURU	AsiaOr+Pacific	13												Subdesenvolupat		Republica
17	NIUE	AsiaOr+Pacific	1			0	100								Subdesenvolupat		Monarquia
18	PALAU	AsiaOr+Pacific	20	6870		9	83								Subdesenvolupat		Republica
19	PAPUA NOVA GUINEA	AsiaOr+Pacific	5772	580	56	11	45	16			63	51	10	30	enDesenvolupament	MMesDefensa	Monarquia
20	SAMOA	AsiaOr+Pacific	184	1860	71	4	100				99	98	6	28	Subdesenvolupat		Monarquia
21	SINGAPUR	AsiaOr+Pacific	4273	24220	79	8		4,1			97	89	5	9	enDesenvolupament	MesSalut	Republica
22	TAILÀNDIA	AsiaOr+Pacific	63694	2540	70	9	99	0,2			95	91	7	16	enDesenvolupament	MMesDefensa	Monarquia
23	TIMOR-LESTE	AsiaOr+Pacific	887	550	56	12	33			6			12	50	Subdesenvolupat		Republica
24	TOKELAU	AsiaOr+Pacific													Subdesenvolupat		Monarquia
25	TONGA	AsiaOr+Pacific	102	1830	72	0	97				99	99	6	24	Subdesenvolupat		Monarquia
26	TUVALU	AsiaOr+Pacific	10			5	88								Subdesenvolupat		Monarquia
27	VANUATU	AsiaOr+Pacific	207	1340	69	6	50						6	31	enDesenvolupament		Republica
28	VIETNAM	AsiaOr+Pacific	83123	550	71	9	41	220		60	94	87	6	20	enDesenvolupament		Republica
29	XINA	AsiaOr+Pacific	1307989	1290	72	4	44	840			95	87	7	13	enDesenvolupament	MesSalut	Republica
30	ANGOLA	AfricaOr+Merid	15490	1030	41	12	30	240			82	54	22	48	enDesenvolupament	MesSalut	Republica
31	BOTSWANA	AfricaOr+Merid	1769	4340	35	10	41	350	90	93	76	82	27	26	enDesenvolupament	MesSalut	Republica
32	BURUNDI	AfricaOr+Merid	7282	90	44	16	36	250		47	67	52	19	45	enDesenvolupament	MesSalut	Republica
33	COMORAS	AfricaOr+Merid	777	530	64	25	23			41	63	49	7	36	enDesenvolupament		Republica
34	ERITREA	AfricaOr+Merid	4232	180	54	21	9	60		62			11	39	enDesenvolupament		Republica
35	ETIOPIA	AfricaOr+Merid	75600	110	48	15	6	1500			49	34	16	41	enDesenvolupament	MesSalut	Republica
36	KENYA	AfricaOr+Merid	33467	460	48	10	48	1200	68	59	78	70	15	39	enDesenvolupament	MMesDefensa	Republica
37	LESOTHO	AfricaOr+Merid	1798	740	35	14	37	320		58	74	90	25	28	enDesenvolupament	MMesDefensa	Monarquia
38	MADAGASCAR	AfricaOr+Merid	18113	300	56	17	33	140	54	49	76	65	12	39	enDesenvolupament	MMesDefensa	Republica

Dades UNICEF 2004 http://www.unicef.org/spanish/infobycountry/index.html

missing patterns

n3=9, x*3=2, r3= 13-2=11

$$\overline{X}_3 = [\overline{X}_{Pob}, \overline{X}_{INB}, \overline{X}_{EV}, \overline{X}_{\%NbP}, \overline{X}_{AlfaH}, \overline{X}_{AlfaD}, \overline{X}_M, \overline{X}_N]$$

$$\bar{X}_3^* = [\bar{X}_{Pob}^*, \bar{X}_{INB}^*, \bar{X}_{EV}^*, \bar{X}_{\%NbP}^*, \bar{X}_{AlphaH}^*, \bar{X}_{AlphaD}^*, \bar{X}_M^*, \bar{X}_N^*]$$

$\hat{\sum}_3$ = *Variances and covariances of full variables*

$\overline{X}_{Pob} = 38120$

$\overline{X}^*_{Pob} = 172062$

$$d^2 = \sum_{j=1}^{J} n_j (\overline{X}_j - \overline{X}_j^*)^T \frac{1}{\hat{\sum}_j} (\overline{X}_j - \overline{X}_j^*) \sim \chi^2_{\sum r_j - K}$$

1	Països		PobTotal	INBzcap	Esper Vida	%NounatsBaixPes	UsInstal Saneja	VIH	%coneixPresHomes	$ConeixPresDones	Alfabet Homes	Alfabet Dones	Mortalitat	Natalitat	Economia	despesaSalut/Defensa	Règim
2	CAMBOYA	AsiaOr+Pacific	13798	320	57	11	16	170		64	85	64	11	31	enDesenvolupament		Monarquia
3	COREA	AsiaOr+Pacific	22384		63	7	59						11	16	enDesenvolupament		Republica
4	FIJI	AsiaOr+Pacific	841	2690	68	10	98	0,6			94	91	6	23	enDesenvolupament	MMesDefensa	Republica
5	FILIPINES	AsiaOr+Pacific	81617	1170	71	20	73	9	59	44	93	93	5	25	enDesenvolupament	MesSalut	Republica
6	ILLES COOK	AsiaOr+Pacific	18			3	100								Subdesenvolupat		Monarquia
7	ILLES MARSHALL	AsiaOr+Pacific	60	2370		12	82								Subdesenvolupat		Republica
8	ILLES SALOMÓ	AsiaOr+Pacific	466	550	63	13	31						7	33	Subdesenvolupat		Monarquia
9	INDONÈSIA	AsiaOr+Pacific	220077	1140	67	9	52	110		23	92	83	7	21	enDesenvolupament	MesSalut	Republica
10	KIRIBATI	AsiaOr+Pacific	97	970		5	39								Subdesenvolupat		Republica
11	LAO, REPÚBLICA DEMOCRÀTICA	AsiaOr+Pacific	5792	390	55	14	24	1,7			77	61	12	35	enDesenvolupament		Republica
12	MALÀSIA	AsiaOr+Pacific	24894	4650	73	9		52			92	85	5	22	enDesenvolupament	MesSalut	Monarquia
13	MICRONÈSIA	AsiaOr+Pacific	110	1990	68	18	28						6	31	Subdesenvolupat		Republica
14	MONGÒLIA	AsiaOr+Pacific	2614	590	65	7	59	0,5		77	98	98	7	22	enDesenvolupament	MesSalut	Republica
15	MYANMAR	AsiaOr+Pacific	50004	220	61	15	73	330			94	86	10	20	enDesenvolupament	MesSalut	Dictadura
16	NAURU	AsiaOr+Pacific	13												Subdesenvolupat		Republica
17	NIUE	AsiaOr+Pacific	1			0	100								Subdesenvolupat		Monarquia
18	PALAU	AsiaOr+Pacific	20	6870		9	83								Subdesenvolupat		Republica
19	PAPUA NOVA GUINEA	AsiaOr+Pacific	5772	580	56	11	45	16			63	51	10	30	enDesenvolupament	MMesDefensa	Monarquia
20	SAMOA	AsiaOr+Pacific	184	1860	71	4	100				99	98	6	28	Subdesenvolupat		Monarquia
21	SINGAPUR	AsiaOr+Pacific	4273	24220	79	8		4,1			97	89	5	9	enDesenvolupament	MesSalut	Republica
22	TAILÀNDIA	AsiaOr+Pacific	63694	2540	70	9	99	0,2			95	91	7	16	enDesenvolupament	MMesDefensa	Monarquia
23	TIMOR-LESTE	AsiaOr+Pacific	887	550	56	12	33			6			12	50	Subdesenvolupat		Republica
24	TOKELAU	AsiaOr+Pacific													Subdesenvolupat		Monarquia
25	TONGA	AsiaOr+Pacific	102	1830	72	0	97				99	99	6	24	Subdesenvolupat		Monarquia
26	TUVALU	AsiaOr+Pacific	10			5	88								Subdesenvolupat		Monarquia
27	VANUATU	AsiaOr+Pacific	207	1340	69	6	50						6	31	enDesenvolupament		Republica
28	VIETNAM	AsiaOr+Pacific	83123	550	71	9	41	220		60	94	87	6	20	enDesenvolupament		Republica
29	XINA	AsiaOr+Pacific	1307989	1290	72	4	44	840			95	87	7	13	enDesenvolupament	MesSalut	Republica
30	ANGOLA	AfricaOr+Merid	15490	1030	41	12	30	240			82	54	22	48	enDesenvolupament	MesSalut	Republica
31	BOTSWANA	AfricaOr+Merid	1769	4340	35	10	41	350	90	93	76	82	27	26	enDesenvolupament	MesSalut	Republica
32	BURUNDI	AfricaOr+Merid	7282	90	44	16	36	250		47	67	52	19	45	enDesenvolupament	MesSalut	Republica
33	COMORAS	AfricaOr+Merid	777	530	64	25	23			41	63	49	7	36	enDesenvolupament		Republica
34	ERITREA	AfricaOr+Merid	4232	180	54	21	9	60		62			11	39	enDesenvolupament		Republica
35	ETIOPIA	AfricaOr+Merid	75600	110	48	15	6	1500			49	34	16	41	enDesenvolupament	MesSalut	Republica
36	KENYA	AfricaOr+Merid	33467	460	48	10	48	1200	68	59	78	70	15	39	enDesenvolupament	MMesDefensa	Republica
37	LESOTHO	AfricaOr+Merid	1798	740	35	14	37	320		58	74	90	25	28	enDesenvolupament	MMesDefensa	Monarquia
38	MADAGASCAR	AfricaOr+Merid	18113	300	56	17	33	140	54	49	76	65	12	39	enDesenvolupament	MMesDefensa	Republica

The Little test in R

- LittleMCAR {BaylorEdPsych}
- **USAGE:** LittleMCAR(x)

  x: dataframe, matrix less than 50 variables

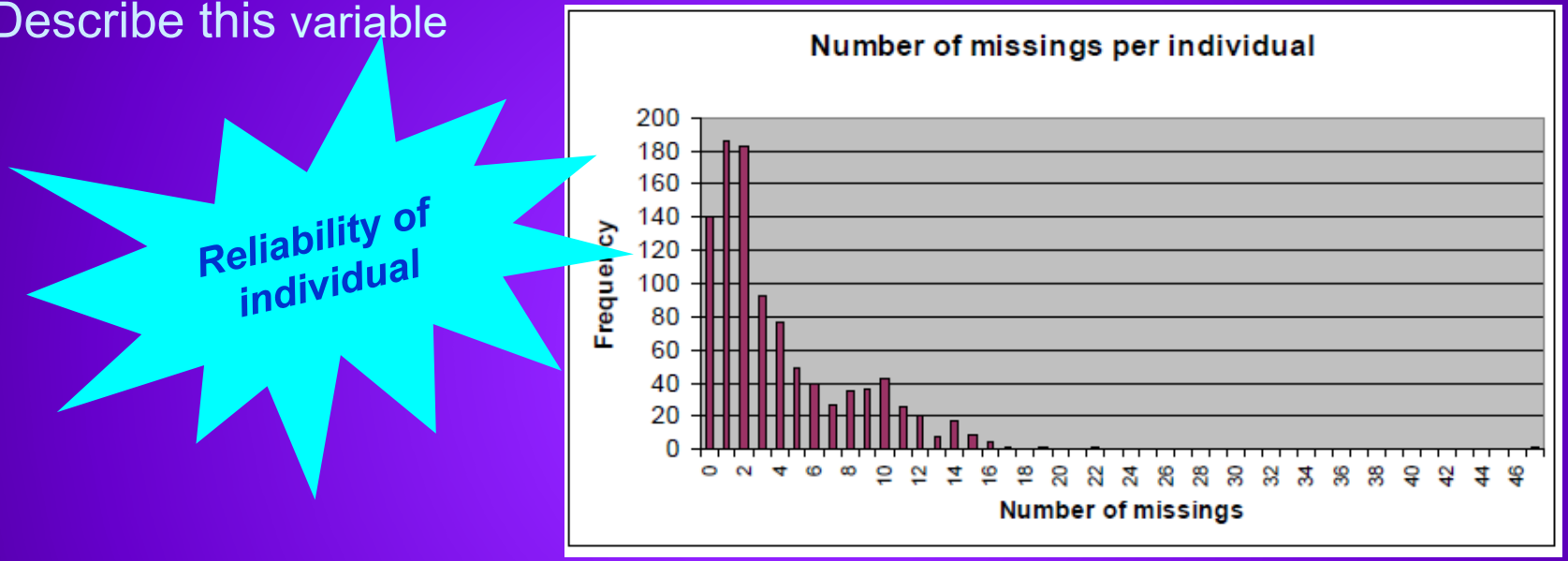
- **Returns:**

chi.square	Chi-square value
df	Degrees of freedom used for chi-square
missing.patterns	Number of missing data patterns
amount.missing	Amount and percent of mssing data
data	The data, organized my missing data patterns

Missing data

(Simple alternative)

- Build new variable counting number of missings per individuals.
- Describe this variable



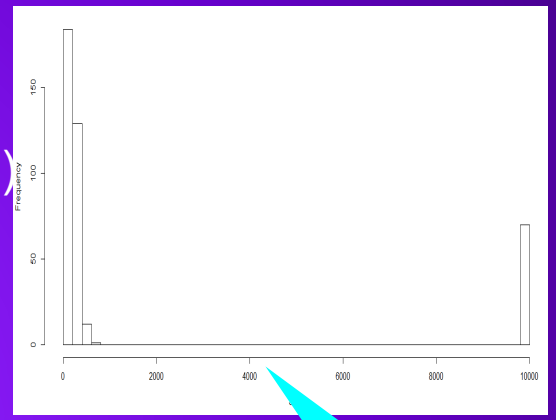
- Count nr of missing per variable and rank variables
  Provides reliability

- Create indicator of missing/non-missing per variable
  and compare both groups of cases

Missing data

(empty cells in data matrix)

- Representation:
  - *,?, “ “, depending on software
  - numerical variables: sometimes codified (0, 99999, -1…)
  - categorical variables: special modality (Ns/Nc, …)



- Standardize missing representation

- Causes of missing data:
  - *voluntary hidden (religion in israel) (always non-random)*
  - *data non-provided*
  - *data non-achieveable*
    - *technical limitations (example anemometers IKE hurrican)*
    - *accessibility (no privileges, sensitive information)*
  - *data lost*
  - *data forced to missing (as a result of correction)*

- *Identification:*
  - *Numerical indicators (stdev…)*

Can be wrongly incorporated in analysis

Missing data treatment

Depends on analysis goals!!!!!

- keep it as a missing: only eventually
  - Can signifficantly reduce the treated observations
- Inputing: Substituting by a useful value (open problem, difficult)
  - Qualitative variable: Substitute by “Unkown<varName>”
  - Stardard way, expert knowledge required
    - use 0
    - use global mean
    - use conditional mean for local groups
    - inputation models (complex)
    - Nearest neighbor (R)
    - Intelligent inputation
    - MIMMI
  - non parametric aproach (montecarlo methods, multiple inputation)
    - special software required
    - technical hypothesis about variable distributions required
    - Final models integration required
      - *Example: French survey, global incomes of household*
- Essential to treat missing BEFORE creating derivated vars.

Missing data treatment

- Missing values frequent in real data
- Imputation before analysis CRITICAL
- Most statistical packages:
  - simple inputation by global mean
  - listwise deletion (dangerous)
- Specific softwares:
  - dedicated to sophisticated inputation methods
  - highly time consuming
  - non-exportable complete data matrices
- Find a trade-of between precision and simplicity

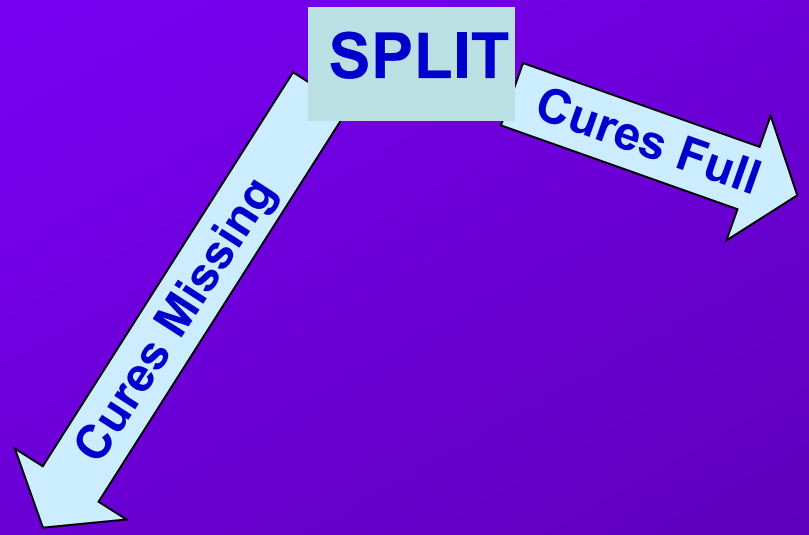
Knn method

C_HISTORI	C_TRACTAI	DATA	Alimentació	Cures d'aparença	Higiene	Vestit: part superior	Vestit: part inferior	Utilització del bany	Bufeta	Intestí	Llit, cadira, cadira de rod
1569,0	84585,0	09/07/2003 0:00	7	7	6	7	7	6	5	5	7
1642,0	74011,0	20/06/2002 0:00	7	7	7	7	7	7	7	7	7
1645,0	84990,0	21/07/2003 0:00	7		6	6	2	6	6	6	3
1666,0	91980,0	09/03/2004 0:00	7	7	7	7	7	6	6	5	7
1694,0	83561,0	03/06/2003 0:00	7	7	7	7	7	7	6	6	7
1754,0	114451,0	03/02/2006 0:00	7	7	6	7	6	6	6	6	7
1858,0	76281,0	26/09/2002 0:00	7		5	7	7	6	5	5	7
1900,0	84368,0	01/07/2003 0:00	6	6	4	4	3	1	6	4	7
1904,0	82443,0	30/04/2003 0:00	4	7	4	6	5	3	2	3	4
1919,0	74098,0	20/06/2002 0:00	7	7	7	7	7	7	6	6	4
1976,0	80110,0	13/02/2003 0:00	7	5	3	4	3	3	5	5	3
2052,0	81175,0	20/03/2003 0:00	7	7	6	7	6	6	6	6	7
2059,0	82951,0	15/05/2003 0:00	1	1	1	1	1	1	1	1	1
2251,0	76399,0	01/10/2002 0:00	5	5	1	1	1	1	6	5	1
2267,0	86796,0	01/10/2003 0:00	7		7	7	7	7	6	6	7
2524,0	76436,0	02/10/2002 0:00	7	7	6	7	6	6	6	6	7
2533,0	81445,0	28/03/2003 0:00	7	7	7	7	7	7	6	6	7
2604,0	75742,0	06/09/2002 0:00	7	7	6	7	7	7	5	6	7
2646,0	84112,0	20/06/2003 0:00	7	7	7	7	7	7	6	6	7
2685,0	79191,0	15/01/2003 0:00	7	7	7	7	7	7	6	6	7
2694,0	78901,0	02/01/2003 0:00	7	7	7	7	7	7	6	6	7
2726,0	74218,0	27/06/2002 0:00	6	6	4	6	6	5	3	5	6
2765,0	79837,0	05/02/2003 0:00	5	5	2	5	2	1	5	5	4

Knn method

Original uncomplete data

C_HISTORI	C_TRACTAI	DATA	Alimentació	Cures d'aparença	Higiene	Vestit: part superior	Vestit: part inferior	Utilització del bany	Bufeta	Intestí	Llit, cadira, cadira de rod
1569,0	84585,0	09/07/2003 0:00	7	7	6	7	7	6	5	5	7
1642,0	74011,0	20/06/2002 0:00	7	7	7	7	7	7	7	7	7
1645,0	84990,0	21/07/2003 0:00	7		6	6	2	6	6	6	3
1666,0	91980,0	09/03/2004 0:00	7	7	7	7	7	6	6	5	7
1694,0	83561,0	03/06/2003 0:00	7	7	7	7	7	7	6	6	7
1754,0	114451,0	03/02/2006 0:00	7	7	6	7	6	6	6	6	7
1858,0	76281,0	26/09/2002 0:00	7		5	7	7	6	5	5	7
1900,0	84368,0	01/07/2003 0:00	6	6	4	4	3	1	6	4	7
1904,0	82443,0	30/04/2003 0:00	4	7	4	6	5	3	2	3	4
1919,0	74098,0	20/06/2002 0:00	7	7	7	7	7	7	6	6	4
1976,0	80110,0	13/02/2003 0:00	7	5	3	4	3	3	5	5	3
2052,0	81175,0	20/03/2003 0:00	7	7	6	7	6	6	6	6	7
2059,0	82951,0	15/05/2003 0:00	1	1	1	1	1	1	1	1	1
2251,0	76399,0	01/10/2002 0:00	5	5	1	1	1	1	6	5	1
2267,0	86796,0	01/10/2003 0:00	7		7	7	7	7	6	6	7
2524,0	76436,0	02/10/2002 0:00	7	7	6	7	6	6	6	6	7
2533,0	81445,0	28/03/2003 0:00	7	7	7	7	7	7	6	6	7
2604,0	75742,0	06/09/2002 0:00	7	7	6	7	7	7	5	6	7
2646,0	84112,0	20/06/2003 0:00	7	7	7	7	7	7	6	6	7
2685,0	79191,0	15/01/2003 0:00	7	7	7	7	7	7	6	6	7
2694,0	78901,0	02/01/2003 0:00	7	7	7	7	7	7	6	6	7
2726,0	74218,0	27/06/2002 0:00	6	6	4	6	6	5	3	5	6
2765,0	79837,0	05/02/2003 0:00	5	5	2	5	2	1	5	5	4



C_HISTORI	C_TRACTAI	DATA	Alimentació	Cures d'aparença	Higiene	Vestit: part superior	Vestit: part inferior	Utilització del bany	Bufeta	Intestí	Llit, cadira, cadira de rod
1569,0	84585,0	09/07/2003 0:00	7	7	6	7	7	6	5	5	7
1642,0	74011,0	20/06/2002 0:00	7	7	7	7	7	7	7	7	7
1666,0	91980,0	09/03/2004 0:00	7	7	7	7	7	6	6	5	7
1694,0	83561,0	03/06/2003 0:00	7	7	7	7	7	7	6	6	7
1754,0	114451,0	03/02/2006 0:00	7	7	6	7	6	6	6	6	7
1900,0	84368,0	01/07/2003 0:00	6	6	4	4	3	1	6	4	7
1904,0	82443,0	30/04/2003 0:00	4	7	4	6	5	3	2	3	4
1919,0	74098,0	20/06/2002 0:00	7	7	7	7	7	7	6	6	4
1976,0	80110,0	13/02/2003 0:00	7	5	3	4	3	3	5	5	3
2052,0	81175,0	20/03/2003 0:00	7	7	6	7	6	6	6	6	7
2059,0	82951,0	15/05/2003 0:00	1	1	1	1	1	1	1	1	1
2251,0	76399,0	01/10/2002 0:00	5	5	1	1	1	1	6	5	1
2524,0	76436,0	02/10/2002 0:00	7	7	6	7	6	6	6	6	7
2533,0	81445,0	28/03/2003 0:00	7	7	7	7	7	7	6	6	7
2604,0	75742,0	06/09/2002 0:00	7	7	6	7	7	7	5	6	7
2646,0	84112,0	20/06/2003 0:00	7	7	7	7	7	7	6	6	7
2685,0	79191,0	15/01/2003 0:00	7	7	7	7	7	7	6	6	7
2694,0	78901,0	02/01/2003 0:00	7	7	7	7	7	7	6	6	7
2726,0	74218,0	27/06/2002 0:00	6	6	4	6	6	5	3	5	6
2765,0	79837,0	05/02/2003 0:00	5	5	2	5	2	1	5	5	4

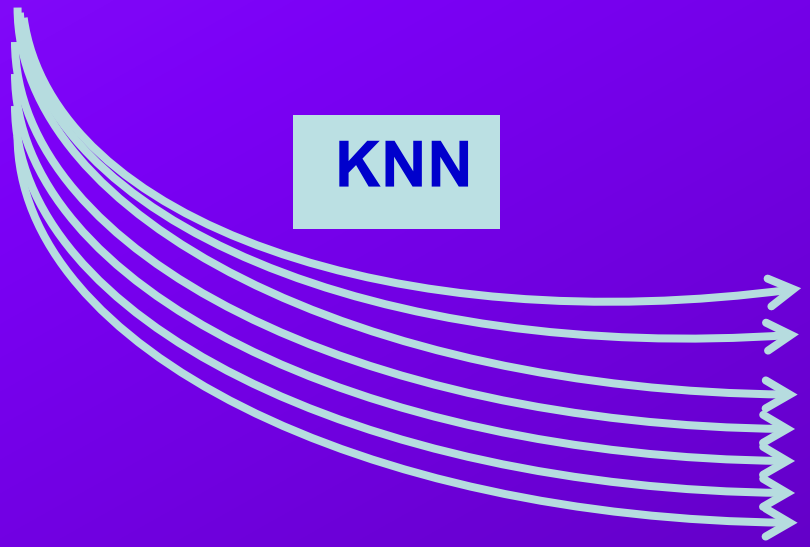
C_HISTORI	C_TRACTAI	DATA	Alimentació	Cures d'aparença	Higiene	Vestit: part superior	Vestit: part inferior	Utilització del bany	Bufeta	Intestí	Llit, cadira, cadira de rod
1645,0	84990,0	21/07/2003 0:00	7		6	6	2	6	6	6	3
1858,0	76281,0	26/09/2002 0:00	7		5	7	7	6	5	5	7
2267,0	86796,0	01/10/2003 0:00	7		7	7	7	7	6	6	7

Knn method

Euclidean distance

Missings in other variables

C_HISTORI	C_TRACTAI	DATA	Alimentació	Cures d'aparenca	Higiene	Vestit: part superior	Vestit: part inferior	Utilització del bany	Bufeta	Intestí	Llit, cadira, cadira de rod
1645,0	84990,0	21/07/2003 0:00	7			6	2	6	6	6	3
1858,0	76281,0	26/09/2002 0:00	7			7	7	6	5	5	7
2267,0	86796,0	01/10/2003 0:00	7			7	7	7	6	6	7



C_HISTORI	C_TRACTAI	DATA	Alimentació	Cures d'aparença	Higiene	Vestit: part superior	Vestit: part inferior	Utilització del bany	Bufeta	Intestí	Llit, cadira, cadira de rod
1569,0	84585,0	09/07/2003 0:00	7		6	7	7	6	5	5	7
1642,0	74011,0	20/06/2002 0:00	7		7	7	7	7	7	7	7
1666,0	91980,0	09/03/2004 0:00	7		7	7	7	6	6	5	7
1694,0	83561,0	03/06/2003 0:00	7		7	7	7	7	6	6	7
1754,0	114451,0	03/02/2006 0:00	7		6	7	6	6	6	6	7
1900,0	84368,0	01/07/2003 0:00	6		4	4	3	1	6	4	7
1904,0	82443,0	30/04/2003 0:00	4		4	6	5	3	2	3	4
1919,0	74098,0	20/06/2002 0:00	7		7	7	7	7	6	6	4
1976,0	80110,0	13/02/2003 0:00	7		3	4	3	3	5	5	3
2052,0	81175,0	20/03/2003 0:00	7		6	7	6	6	6	6	7
2059,0	82951,0	15/05/2003 0:00	1		1	1	1	1	1	1	1
2251,0	76399,0	01/10/2002 0:00	5		1	1	1	1	6	5	1
2524,0	76436,0	02/10/2002 0:00	7		6	7	6	6	6	6	7
2533,0	81445,0	28/03/2003 0:00	7		7	7	7	7	6	6	7
2604,0	75742,0	06/09/2002 0:00	7		6	7	7	7	5	6	7
2646,0	84112,0	20/06/2003 0:00	7		7	7	7	7	6	6	7
2685,0	79191,0	15/01/2003 0:00	7		7	7	7	7	6	6	7
2694,0	78901,0	02/01/2003 0:00	7		7	7	7	7	6	6	7
2726,0	74218,0	27/06/2002 0:00	6		4	6	6	5	3	5	6
2765,0	79837,0	05/02/2003 0:00	5		2	5	2	1	5	5	4

Knn method

C_HISTORI	C_TRACTA	DATA	Alimentació	Cures d'aparença	Higiene	Vestit: part superior	Vestit: part inferior	Utilització del bany	Bufeta	Intestí	Llit, cadira, cadira de rod
1645,0	84990,0	21/07/2003 0:00	7			6	7	6	6	6	7
1858,0	76281,0	26/09/2002 0:00	7			7	7	6	5	5	7
2267,0	86796,0	01/10/2003 0:00	7			7	7	7	6	6	7

7

KNN

C_HISTORI	C_TRACTA	DATA	Alimentació	Cures d'aparença	Higiene	Vestit: part superior	Vestit: part inferior	Utilització del bany	Bufeta	Intestí	Llit, cadira, cadira de rod
1569,0	84585,0	09/07/2003 0:00	7		6	7	7	6	5	5	7
1642,0	74011,0	20/06/2002 0:00	7		7	7	7	7	7	7	7
1666,0	91980,0	09/03/2004 0:00	7		7	7	7	6	6	5	7
1694,0	83561,0	03/06/2003 0:									
1754,0	114451,0	03/02/2006 0:00	7		6	7	6	6	6	6	7
1900,0	84368,0	01/07/2003 0:00	6		4	4	3	1	6	4	7
1904,0	82443,0	30/04/2003 0:00	4		4	6	5	3	2	3	4
1919,0	74098,0	20/06/2002 0:00	7		7	7	7	7	6	6	4
1976,0	80110,0	13/02/2003 0:00	7		3	4	3	3	5	5	3
2052,0	81175,0	20/03/2003 0:00	7		6	7	6	6	6	6	7
2059,0	82951,0	15/05/2003 0:00	1		1	1	1	1	1	1	1
2251,0	76399,0	01/10/2002 0:00	5		1	1	1	1	6	5	1
2524,0	76436,0	02/10/2002 0:00	7		6	7	6	6	6	6	7
2533,0	81445,0	28/03/2003 0:00	7		7	7	7	7	6	6	7
2604,0	75742,0	06/09/2002 0:00	7		6	7	7	7	5	6	7
2646,0	84112,0	20/06/2003 0:00	7		7	7	7	7	6	6	7
2685,0	79191,0	15/01/2003 0:00	7		7	7	7	7	6	6	7
2694,0	78901,0	02/01/2003 0:00	7		7	7	7	7	6	6	7
2726,0	74218,0	27/06/2002 0:00	6		4	6	6	5	3	5	6
2765,0	79837,0	05/02/2003 0:00	5		2	5	2	1	5	5	4

Mixed Intelligent-Multivariate Missing Imputation
The MIMMI method *[Gibert 2013]*

- Select a small number of relevant variables
  (whith small ratio of missing data)
- Use intelligent inputation on that reduced data matrix
  (expert-based inputation, vertical or horitzontal)
- Multivariate clustering using the imputed variables
- Determine a partition of the data
- Inpute the missing data of the remaining variables
  (use mean local to the group of every individual (conditional means)

Example OMS

Trade-off
Accuracy/required time

Gibert, K (2013) Mixed Intelligent-Multivariate Missing Imputation, Int'l Journal of Computer Mathematics 91(1):85-96

MIMMI *Method [IJCM Gibert 2013]*

MIMMI method on WHO-AIMS database

Selection of 16 core variables

WHO-AIMSname	Meaning	KLASSname
polplanr	Presence of policy or plan	polplanr
legisl	Presence of legislation	Legisl
d1f5i5rec	Affordability of anitpsychotic medicine	d1f5i5rec(antipsych)
d1f5i6rec	Affordability of anitdepressant medicine	d1f5i6rec(antidepr)
D2F1I2	Oragnization of services	D2f1i2(orgServices)
cbusrate	Community based inpatient units per 100,000 population	cbusrate
mhrate	mental hospitals per 100 000 population	mhrate
outpfrate	outpatient facilities per 100 000 population	outpfrate
daytrfrate	day treatment facilities per 100 000	daytrfrate
D4F1I11	psychiatrists per 100 000	d4f1i11(psychi)
D4F1I12	other doctors per 100 000	D4F1I12(doctors)
D4F1I13	nurses per 100 000	D4F1I13(nurses)
D4F1I14	psychologists per 100 000	D4F1I14(psycho)
D4F1I15	Social workers per 100 000	D4F1I15(socWorK)
d3f1i3	availability of treatment and assessment manuals	d3f1i3(manuals)
d5f2i51	formal collaborative relationship with department of primary care	d5f2i51(relprimcare)
D6F1I1	formally defined min data set	D6F1I1(mindataset)

MIMMI *Method* *[IJCM Gibert 2013]*

- Selection of 16 core variables:
  - Characteristic information of the whole 6 domains
  - Related with decisional variables (composite indicators)
  - Low tax of missing data

KLASSname	nMissing
outpfrate	1
D4F1I12(Other)	1
D4F1I13(nurses)	1
D4F1I14(psycho)	2
D4F1I15(socWorK)	1
d3f1i3(manuals)	1
d5f2i51(relprimcare)	1

Complex process

time consuming

applicable in real projects with experts

- Total of 8 missing cells to be inputed

- Intelligent inputation of 8 missing values

MIMMI *Method [IJCM Gibert 2013]*

Intelligent inputation of 8 missing values

Country	Missing variable and imputed value
South Africa	outpfrate = 2.0
China	D4F1I12(Other) = 1.20
India	D4F1I13(nurses) = 0.15
China	D4F1I14(psycho) = 0.16
Paraguay	D4F1I14(psycho) = 1.4
Nepal	D4F1I15(socWorK)= 0.15
Moldova	d3f1i3(manuals) = B
Azerbaijan	d5f2i51(relprimcare) = N

MIMMI *Method* *[IJCM Gibert 2013]*

Intelligent inputation of 8 missing values

Country	Missing variable and imputed value
South Africa	outpfrate = 2.0
China	D4F1I12(Other) = 1.20
India	D4F1I13(nurses) = 0.15
China	D4F1I14(psycho) = 0.16
Paraguay	D4F1I14(psycho) = 1.4
Nepal	D4F1I15(socWorK)= 0.15
Moldova	d3f1i3(manuals) = B
Azerbaijan	d5f2i51(relprimcare) = N

Vertical Imputation

Rate of "other medical doctors" in China
Very institutional system.
Higher than other Asian countries.
Between Vietnam and Thailand.

Two Strategies

MIMMI *Method [IJCM Gibert 2013]*

Intelligent inputation of 8 missing values

Country	Missing variable and imputed value
South Africa	outpfrate = 2.0
China	D4F1I12(Other) = 1.20
India	D4F1I13(nurses) = 0.15
China	D4F1I14(psycho) = 0.16
Paraguay	D4F1I14(psycho) = 1.4
Nepal	D4F1I15(socWorK)= 0.15
Moldova	d3f1i3(manuals) = B
Azerbaijan	d5f2i51(relprimcare) = N

Horizontal Imputation

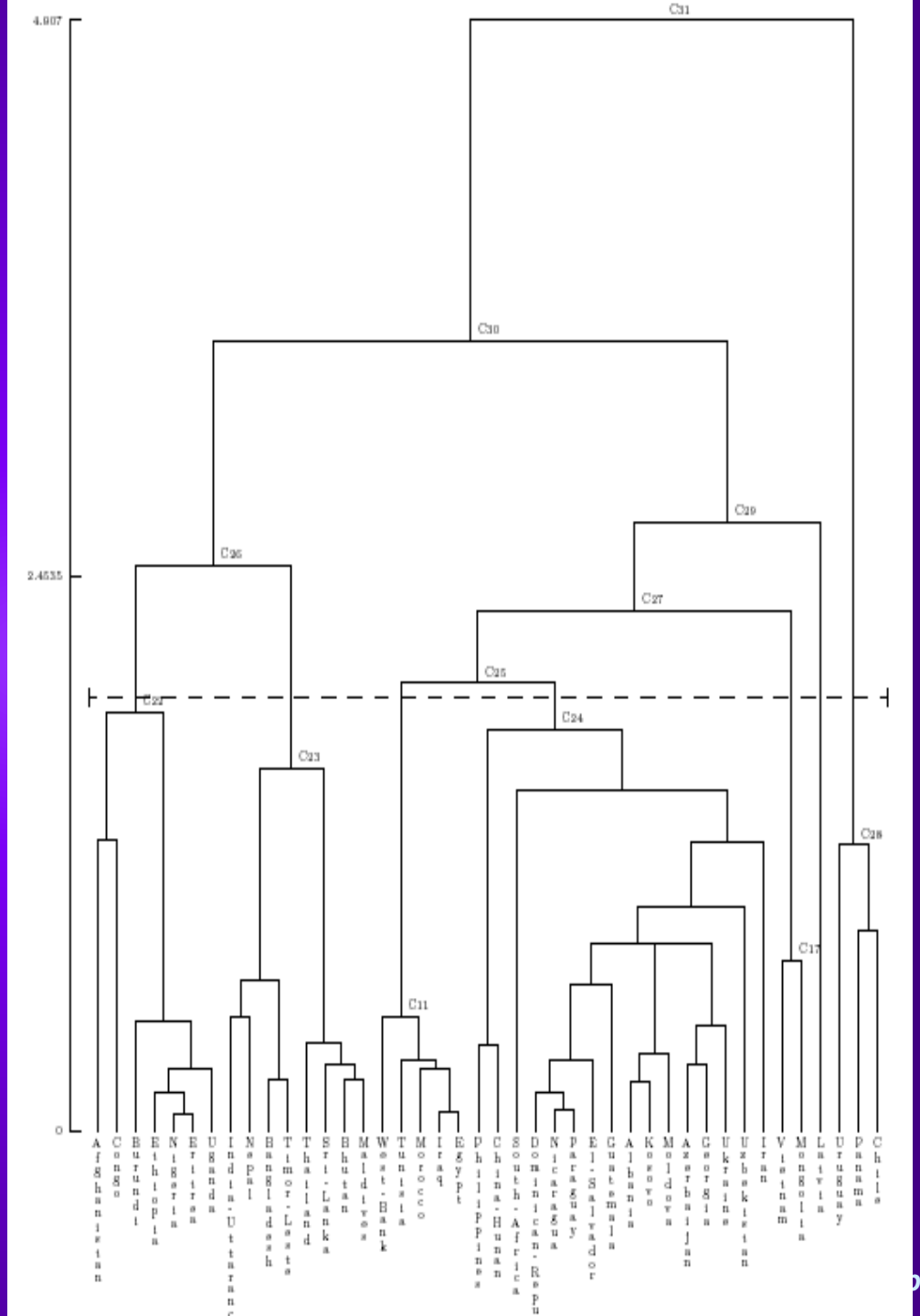
India-Uttaranchal (nurses rate)
High where hospitals are.
India do not has mental hospital

Choose a low value

Two Strategies

MIMMI *Method*

- Inputation:
  - Complete the 42x16 data matrix
  - Clustering the full matrix
    - *Hierarchical*
    - *Ward criterion*
    - *Gibert's mixed metrics*
    - *[Gibert 96]*
  - Determine the classes (7)
  - Find partition



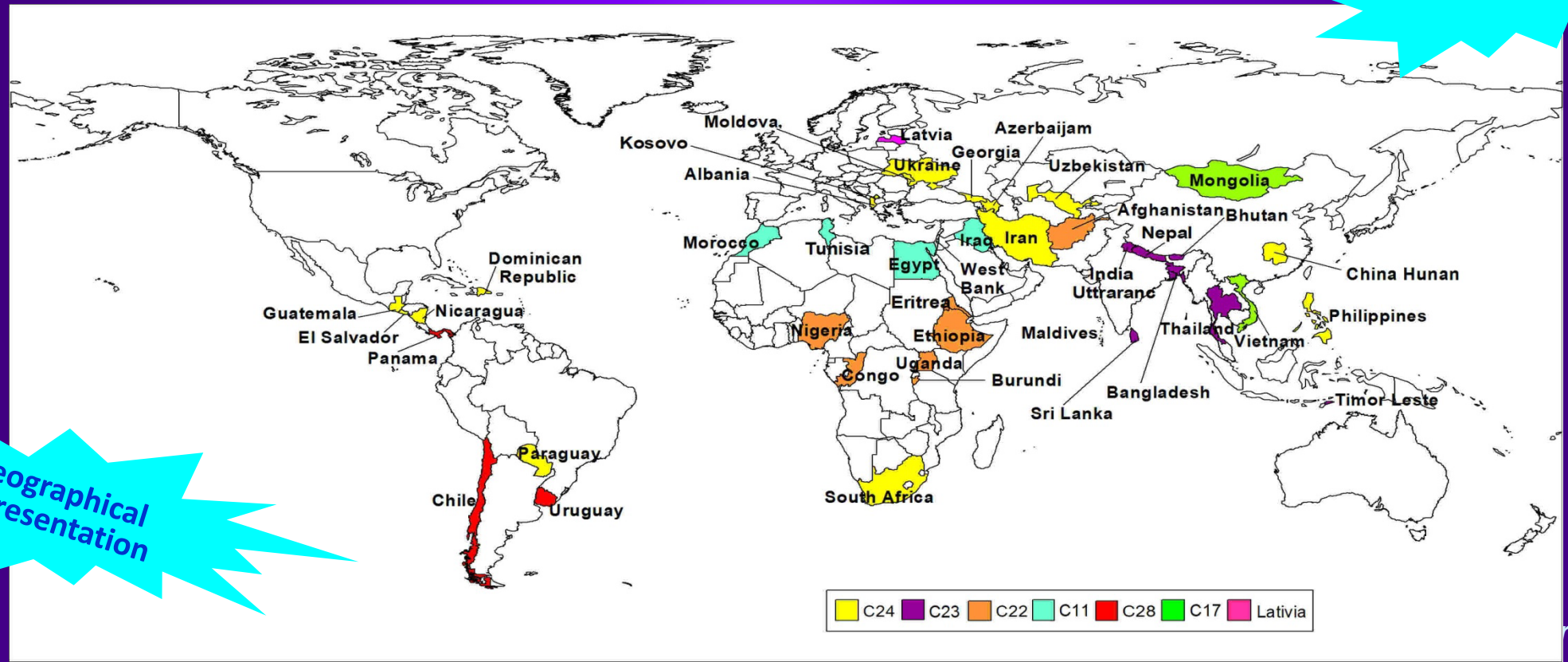
- *Gibert, K., and Cortés, U. (1997). "Weighing quantitative and qualitative variables in clustering methods." Mathware and Soft Computing, 4(3), 251-266.*

MIMMI *Method [IJCM Gibert 2013]*

Seven classes recommended

- **C22:** Afghanistan, Burundi, Congo, Eritrea, Ethiopia, Nigeria, Uganda
- **C24:** Albania, Azerbaijan, China-Hunan, Dominican-Repu, El-Salvador, Georgia, Guatemala, Iran, Kosovo, Moldova, Nicaragua, Paraguay, Philippines, South-Africa, Ukraine, Uzbekistan
- **C23:** Bangladeh, Bhutan, India-Uttaranc, Maldives, Nepal, Sri-Lanka, Thailand, Timor-Leste
- **C28:** Chile, Panama, Uruguay
- **C11:** Egypt, Iraq, Morocco, Tunisia, West-Bank

Burundi



Geographical representation

MIMMI *Method [IJCM Gibert 2013]*

Local class means of numerical variables among the 256 variables

	CLASSE	C22	C24	C23	C28	C11	Latvia	C17
VARIABLE	N = 42	$n_c = 7$	$n_c = 16$	$n_c = 8$	$n_c = 3$	$n_c = 5$	$n_c = 1$	$n_c = 2$
totprofinh	$\bar{X}$	1.28	13.5507	5.1017	21.15	4.53	47.23	8.775
	S	0.9711	10.0276	6.0099	7.5041	2.6071		7.3468
	N*	0	2	2	0	1	0	0
treatpre	$\bar{X}$	192.22	1219.4614	59.7	1037.8201	547.0175	3490.75	1251.71
	S	121.1716	1447.8447	39.4424	519.8082	550.17	?	?
	N*	3	3	6	1	1	0	1
lundpararectrail	$\bar{X}$	1.09	0.9	0.49	1.2	1.1567	0.19	0.76
	S	0.7916	0.5622	?	0.0666	0.8864	?	?
	N*	3	7	7	1	2	0	1
concarewor	$\bar{X}$	0.0314	0.0856	0.0197	0.0259	0.624	0.1991	0.1313
	S	0.0083	0.0196	?	0.0057			
	N*	5	10	7	1	4	0	1
usmhexperca	$\bar{X}$	0.2646	0.4961	0.2466	2.7995	0.4102	10.172	0.256
	S	0.5312	0.5059	0.2915	2.5926	0.2307		
	N*	0	1	1	0	1	0	1
d1f5i2exmhos	$\bar{X}$	0.7783	0.7954	0.7463	0.4963	0.5768	0.804	0.636
	S	0.2019	0.2121	0.1582	0.2051	0.1106		?
	N*	1	1	4	0	1	0	1

MIMMI *Method* *[IJCM Gibert 2013]*

Complex process
highly time consuming
applicable in real projects

- Horizontal inputation:
  use the value of other variables of the same individual as predictors of the missing value.

  inputing 0 in the income of 4th person if the household has only 1,2 or 3 persons

- Vertical inputation:
  use the value of the same variable in other similar individuals

  use the mean of the salary of 4rt persons over 18 years old if the household hasmore than 4per

Mixed Intelligent-Multivariate Missing Imputation
The MIMMI method *[Gibert 2013]*

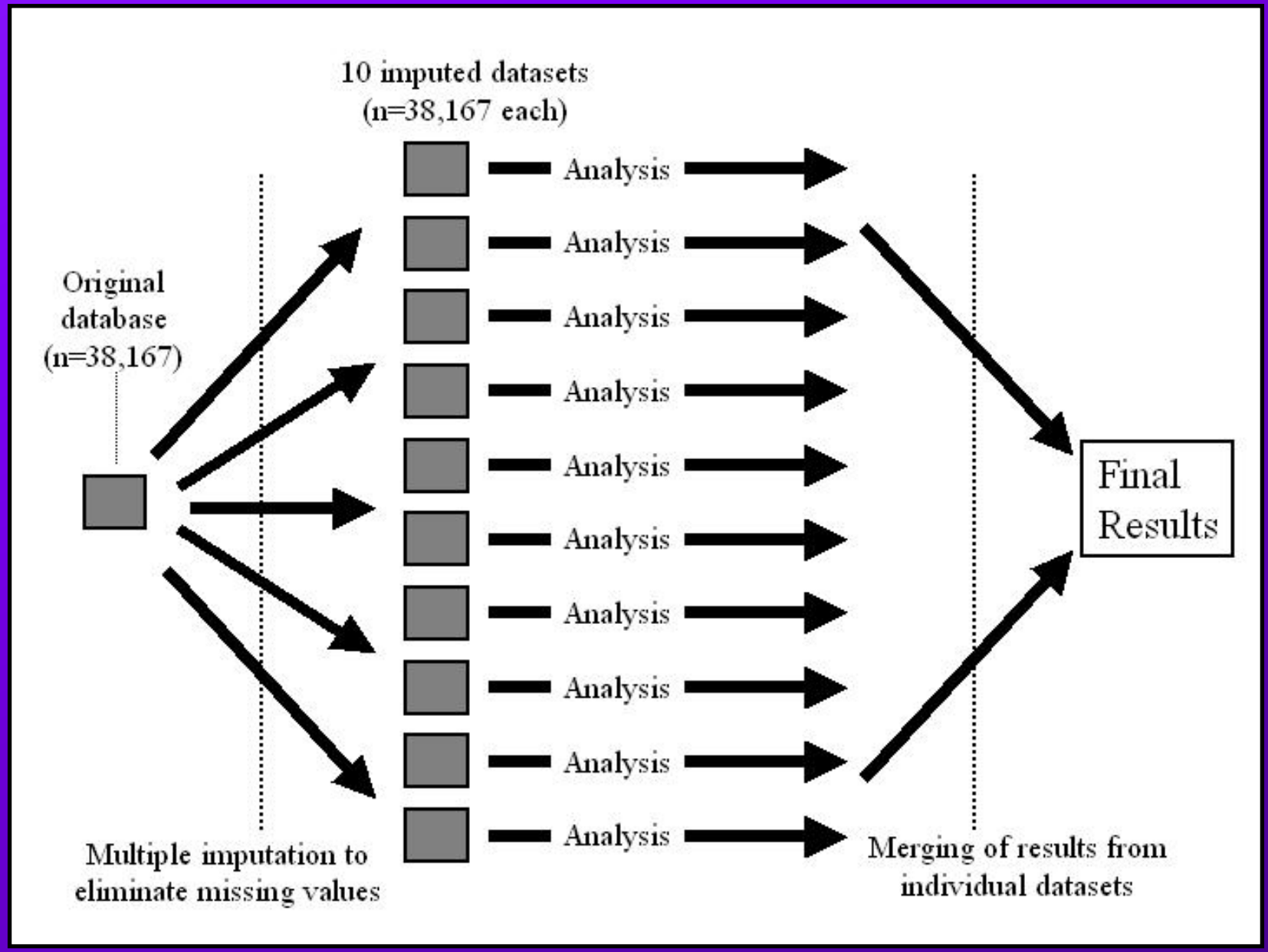
- Select a small number of relevant variables
  (whith small ratio of missing data)
- Use intelligent inputation on that reduced data matrix
  (expert-based inputation, vertical or horitzontal)
- Multivariate clustering using the imputed variables
- Determine a partition of the data
- Inpute the missing data of the remaining variables
  (use mean local to the group of every individual (conditional means)

Example OMS

Trade-off
Accuracy/required time

Gibert, K (2013) Mixed Intelligent-Multivariate Missing Imputation, Int'l Journal of Computer Mathematics 91(1):85-96

MICE



MICE

The overall *estimate of your parameter* (Q-bar) is its mean across the m imputations

$$\bar{Q} = m^{-1} \sum \hat{Q}^{(\ell)}$$

The *within-imputation variance* (U-bar) of the Q parameter is the mean of the variances across the m imputations

$$\bar{U} = m^{-1} \sum U^{(\ell)}$$

The *between-imputation variance* (B) of the Q parameter is standard deviation of Q across the m imputations

$$B = (m-1)^{-1} \sum (\hat{Q}^{(\ell)} - \bar{Q})^2$$

The *total variance* of Q is a function of U-bar and B. This total variance is used to calculate the standard error used for test statistics

$$T = (1 + m^{-1})B + \bar{U}$$

$$(\bar{Q} - Q)/\sqrt{T} \sim t_\nu$$

The *degrees of freedom* (v) are adjusted for the amount of information lost to missing data

$$\nu = (m-1)\left[1 + \frac{\bar{U}}{(1+m^{-1})B}\right]^2$$

MICE

- MICE method is very flexible – but demands thought when creating the imputation model
- Strongly recommend mastering the `eq()`, `passive()` and `substitute()` options
- Can deal with interactions using `passive()`
- Choice of *m* is important
  - may need to be (much) larger than 5
  - See Royston (2004, SJ 4:227-41) for discussion
- available in MICE Rpackage

Easy for MAP-REDUCE strategies

WARNING reproducibility

WARNING final model consensus

Royston, P. (2004). Multiple imputation of missing values. *Stata journal*, *4*(3), 227-41.

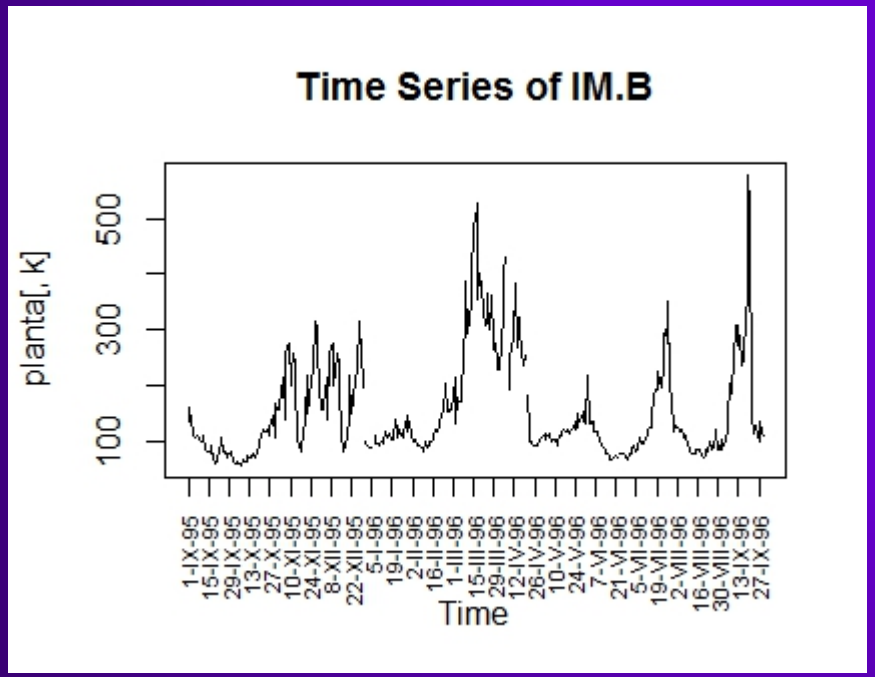
Interpolation

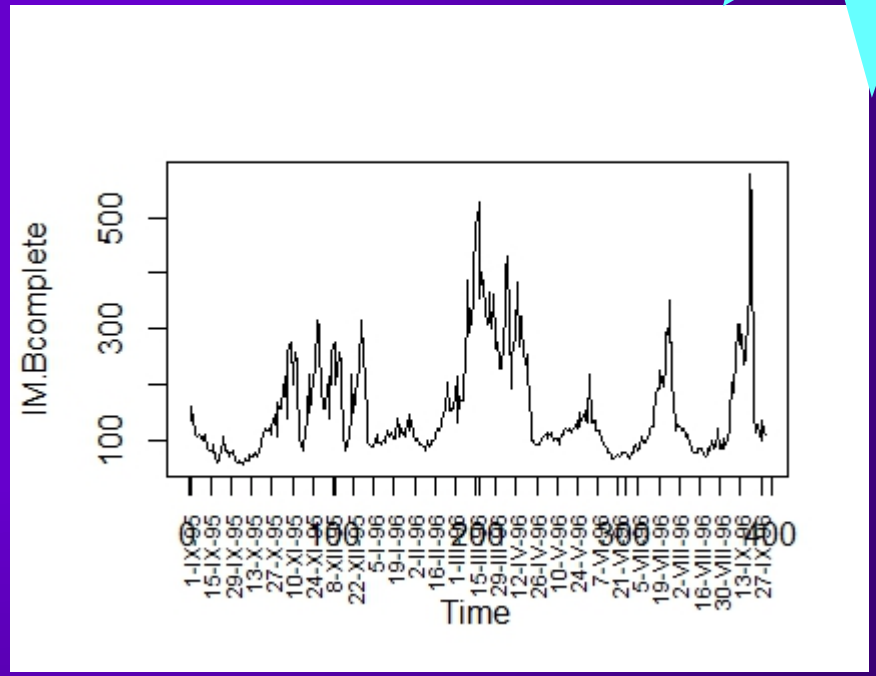
Usefull for time-series of numerical variables

- Linear assumption between observed points

  (assume monotonic behaviour between observations)

ALTERNATIVE
Assume constant between Measurements (slow dynamics)





IM.B: Mass index of mixed liquor (na.approx {zoo})

Preprocessing

Data cleaning
Data preparation
Data preprocessing

- Formatting issues, building software context
- Determining working matrix, Filtering
- Identification and treatment of missing data
- Identification and treatment of outliers
- Identification and treatment of errors *(correct when possible)*
- Feature selection/extraction, dimensionality reduction
- Instance selection
- Data transformation
- Derivation of new variables

Outlier

- Rare observation (presumed out of range)
- Multivariate vs univariate outlier

BIAS

Types of outliers:

- Mistake (Transcription Error or Measurement Error)
  - A person 560 years old
  - FIRST VERIFY *If possible correct*.
    If not, substitute by missing
- Informative point
  - A single informative point of a missing part of the population
  - *Complete the sample*
    when impossible, restrict scope of analysis
- Extreme value of the population
  - Very old person, 99 years old
  - *Keep*
- Value of another population
  - One swedish in the middle of cannibal tribu, measuring height)
  - *Treat apart. CLEARLY REPORT ABOUT IT*
- Missing code
  - *Substitute by missing or inpute*

Care with suppressions

The danger of suppressions

In 1985 British scientists reported a hole in the ozone layer of the earth's atmosphere over the South Pole. This is disturbing, since ozone protects us from cancer-causing ultraviolet radiation. The British report was at first disredarded, since it was based on ground instruments looking up. More comprehensive observations from satellite instruments looking down had shown nothing unusual. Then examination of the satellite data revealed that the South Pole ozone readings were so low that the computer software used o analyze the data had automatically syppressed these values as erroneous outliers. Readings dating back to 1979 were reanalized and showed a large and growing hole in the ozone layer that is unexplained and possibily dangerous Computers analyzing large volumnes of data are often programmed to suppress outliers as protection againts errors in the data. As the example of the hole in the ozone layer illustrated, suppressing an outliers without inv estigating it can kleep valuable information out of the

Moore, McCabe, Introduction to the practice of Statistics, 5th Edition,

From the paper of John Gleick in New York Times, Jul 1985

http://www.nytimes.com/1986/07/29/science/hole-in-ozone-over-south-pole-worries-scientists.html?pagewanted=all

Outlier detection

- Specific statistical Tests:
  Depend on the software

  Usually for specific distributions

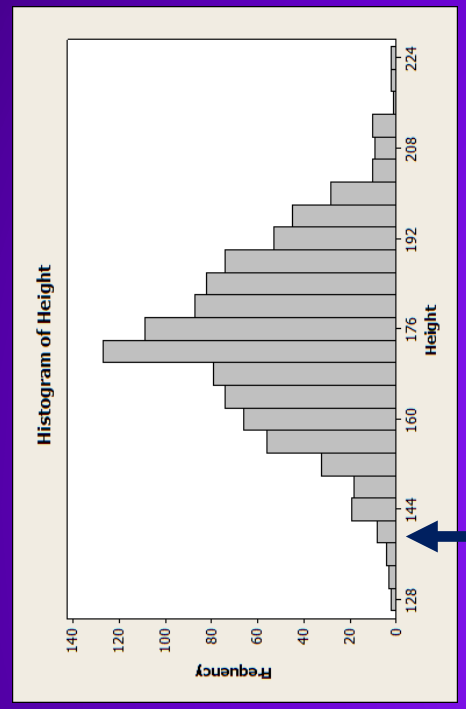
- Graphical representation of the distribution of data

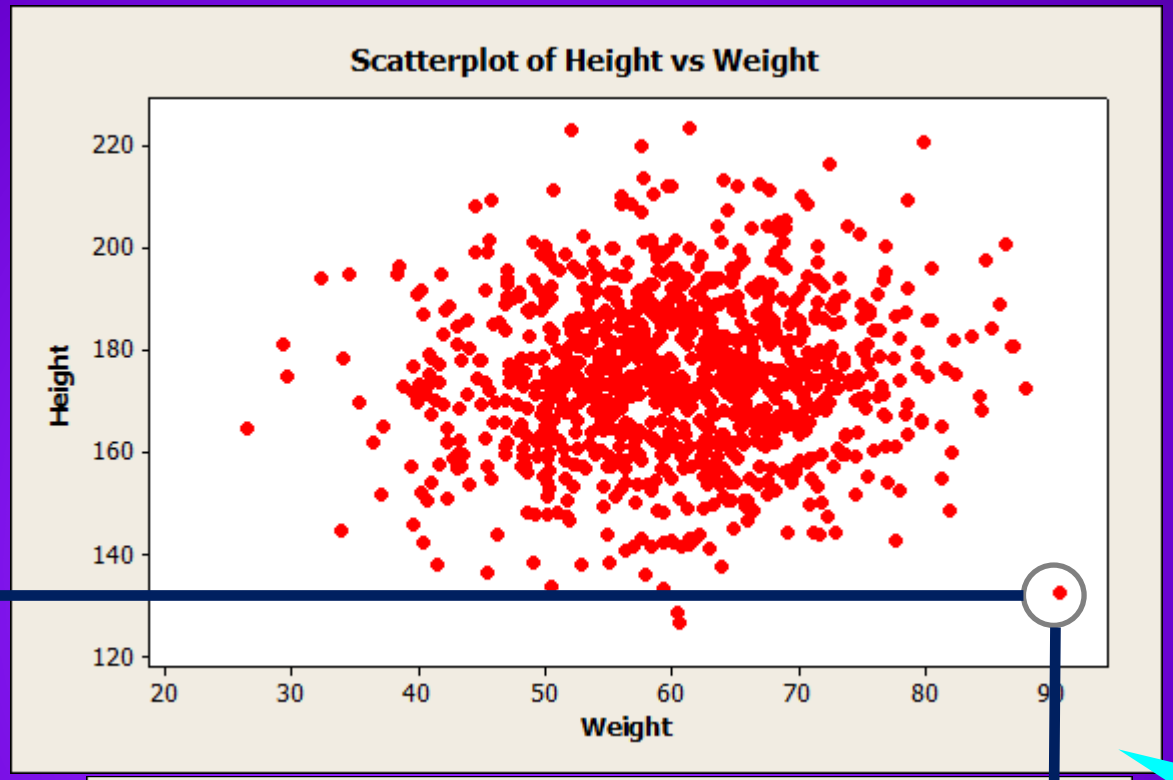
  Univariate (histogram or boxplot)

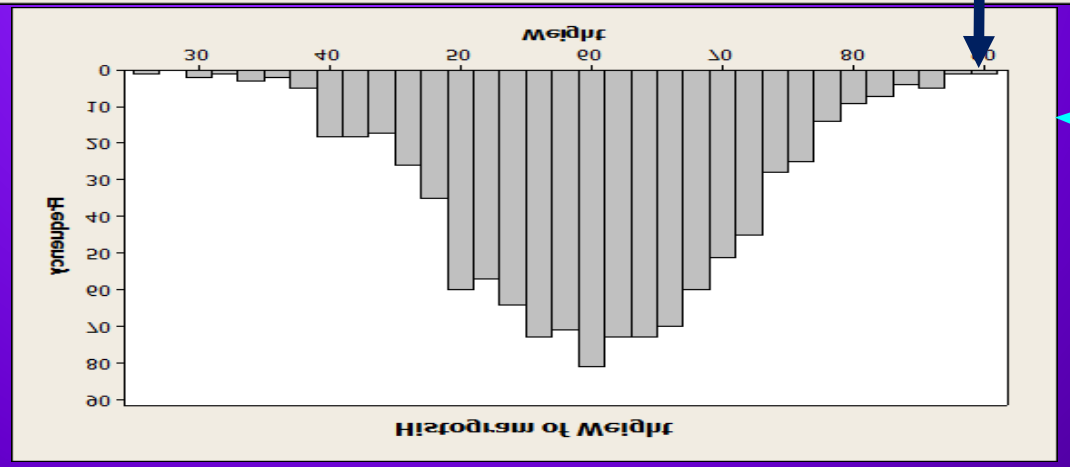
  Bivariate (plots)

  Clustering for multivariate outliers (singletons)

Dimensionality of outliers: Bivariate Outlier



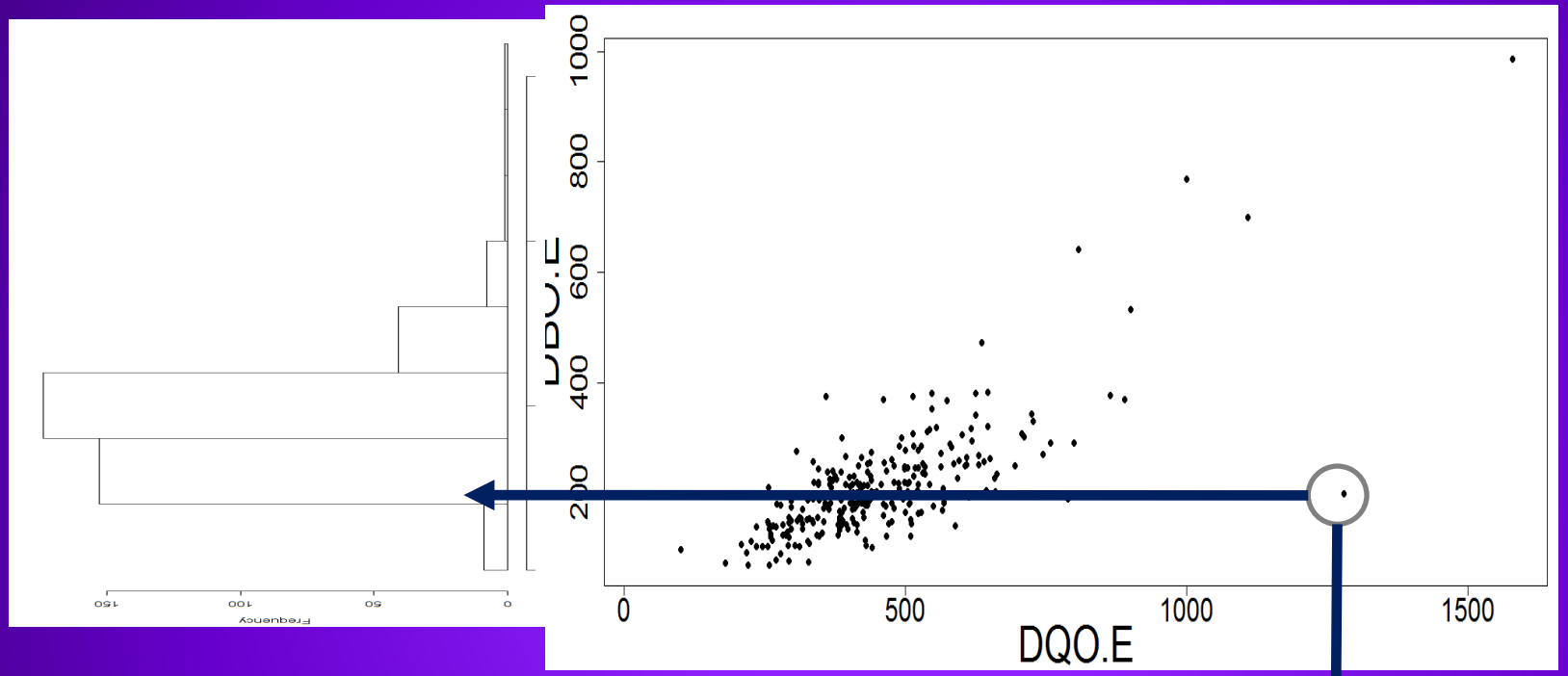




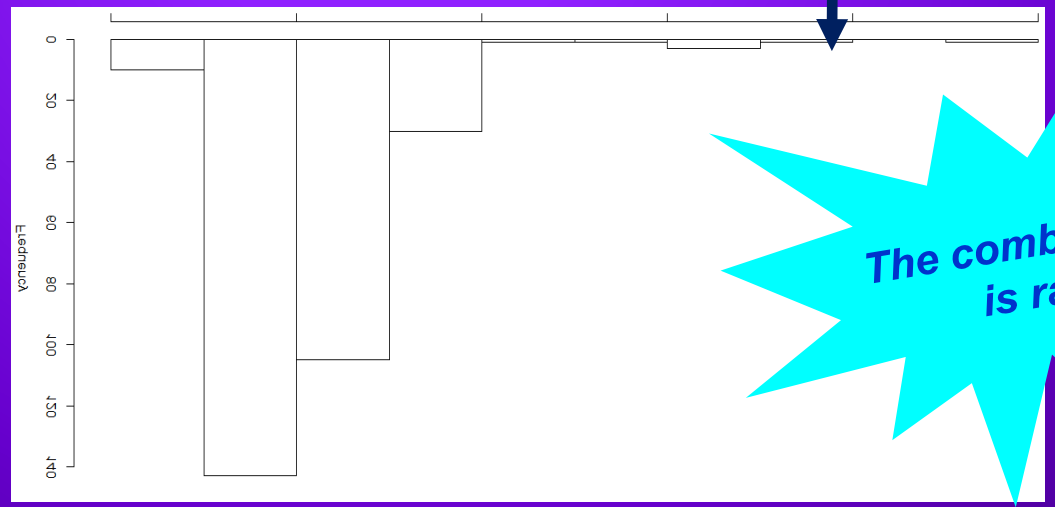
A person with 90Kg and 1,32 m

The combination is rare

Dimensionality of outliers: Bivariate Outlier

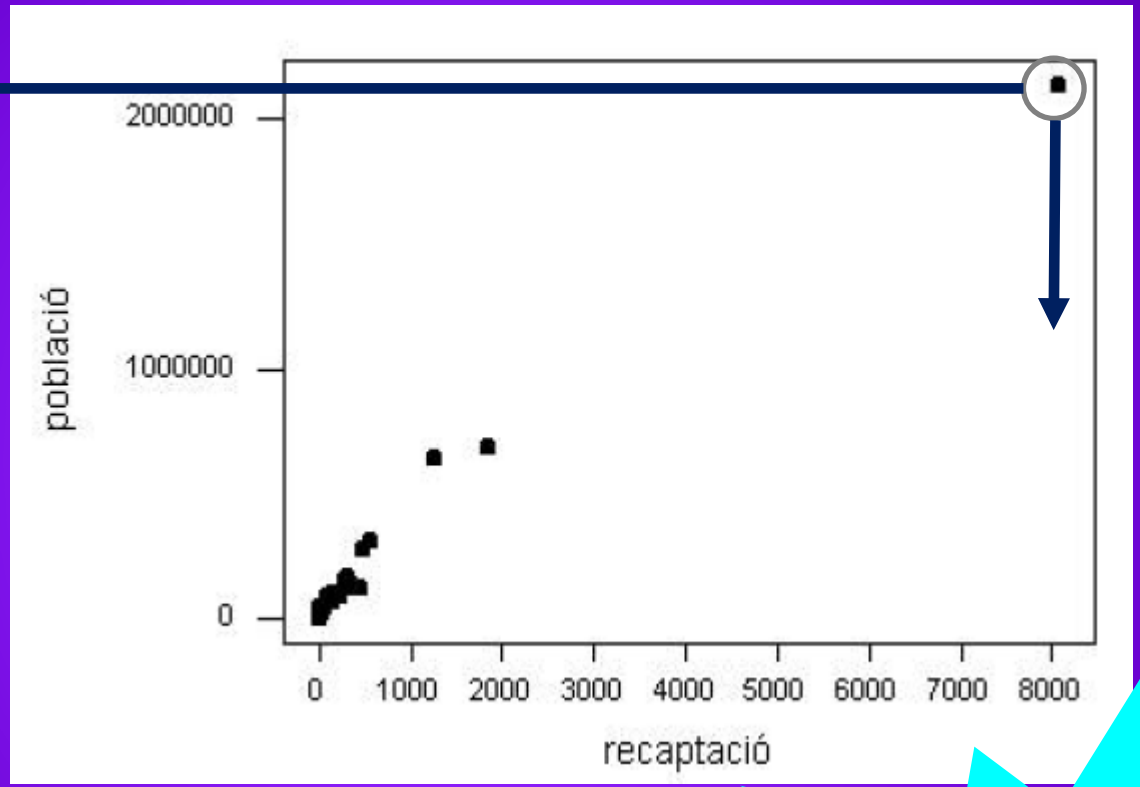


A day with
DQO.E =1279 mg/l
DBO.E = 198 mg/l



The combination is rare

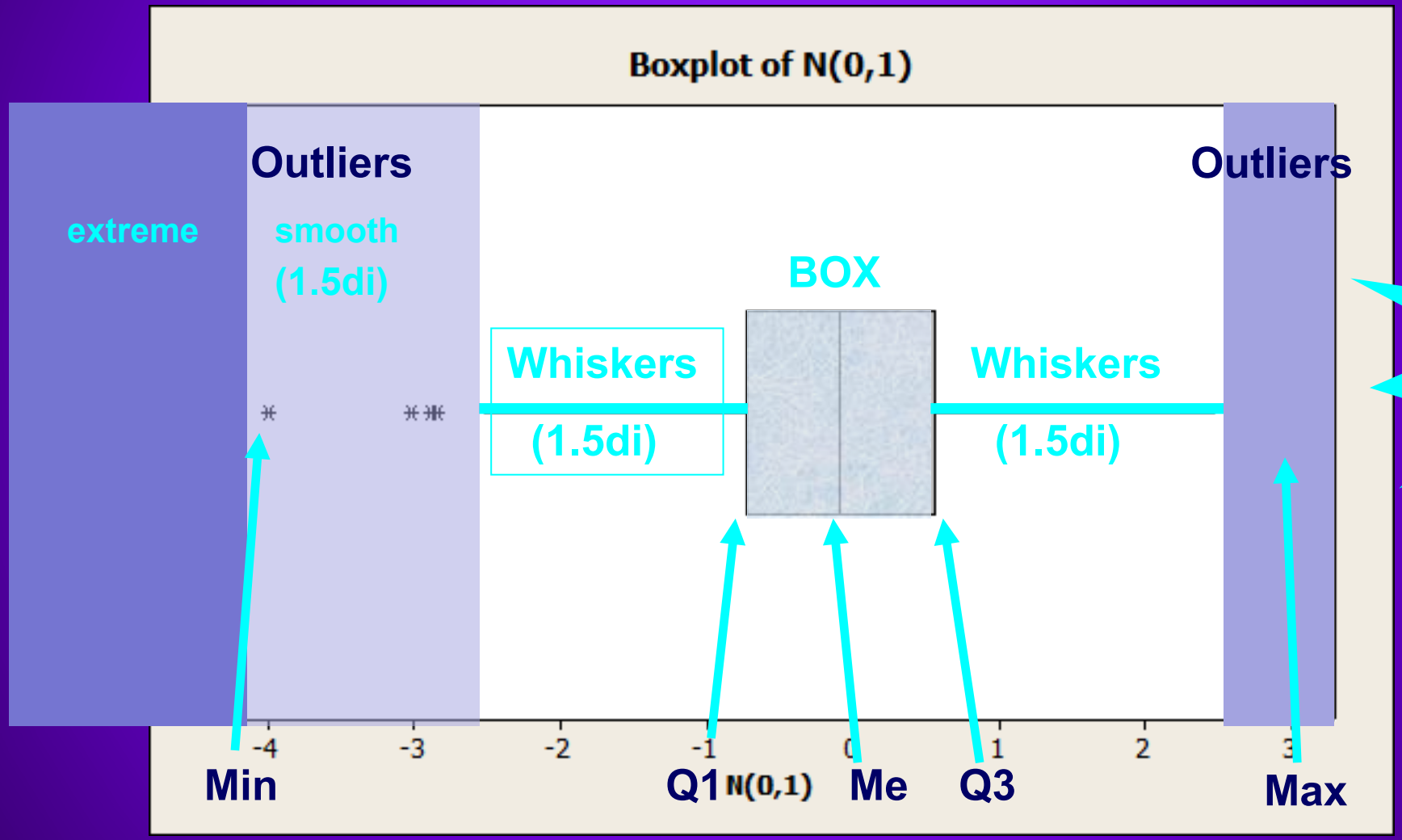
Dimensionality of outliers
Univariate Outlier



Aligned with global model!!!!

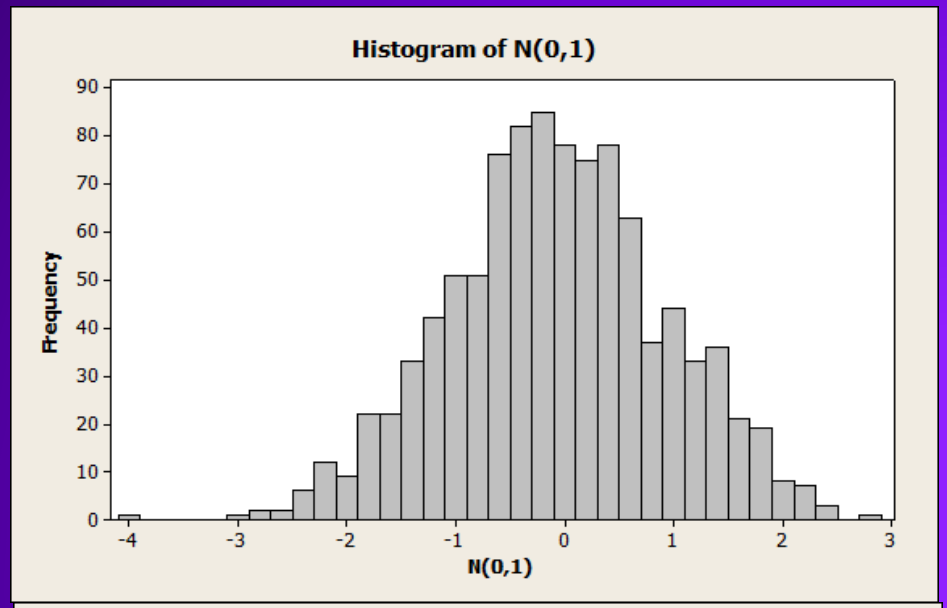
Boxplot *[Tukey 1956]*

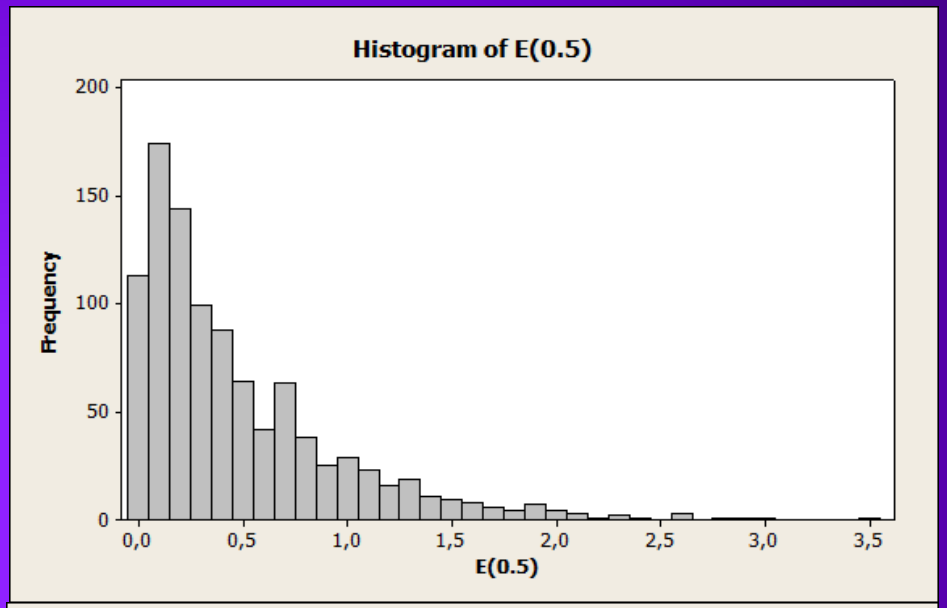
Symbolic representation of empirical distribution

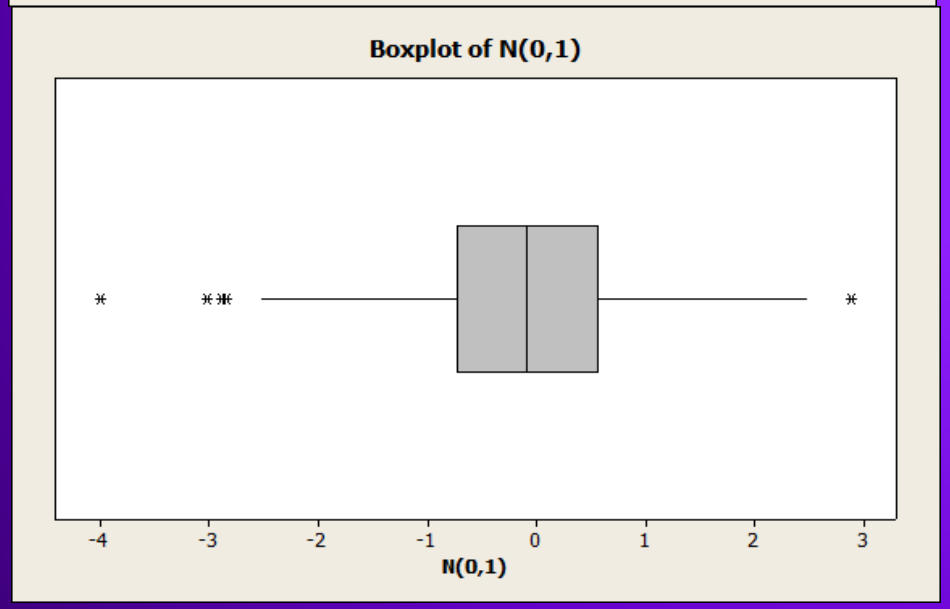


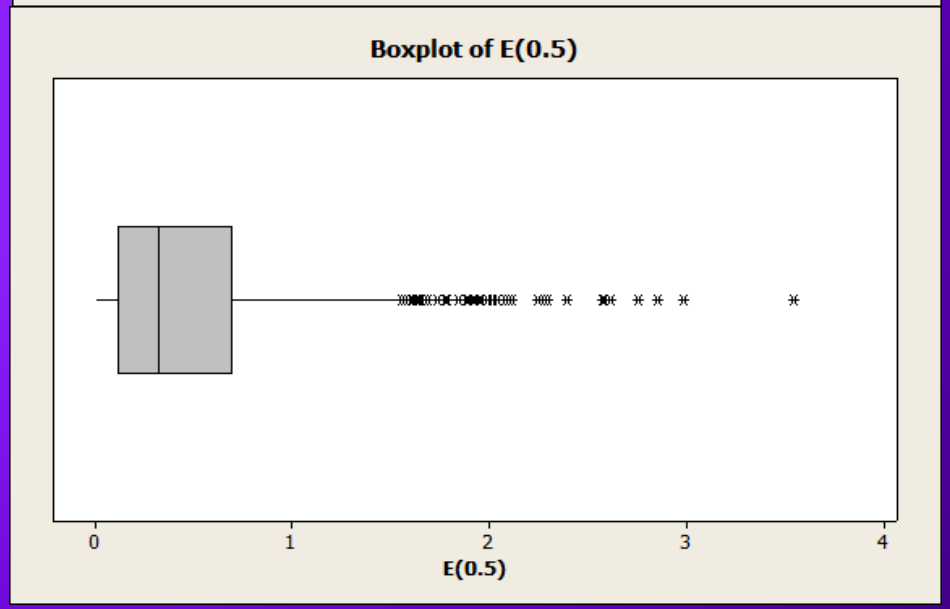
Assumes
Symetric distribution

Boxplot [Tukey 1956]









Preprocessing

Data cleaning
Data preparation
Data preprocessing

- Formatting issues, building software context
- Determining working matrix, Filtering
- Identification and treatment of missing data
- Identification and treatment of outliers
- Identification and treatment of errors *(correct when possible)*
- Feature selection/extraction, dimensionality reduction
- Instance selection
- Data transformation
- Derivation of new variables

Instance selection

Evaluation of representative instances in a dataset

- Elimination of irrelevant instances
- Sampling
- Resampling

Reparing unbalanced datasets when required

- oversampling
- undersampling

Preprocessing

Data cleaning
Data preparation
Data preprocessing

- Formatting issues, building software context
- Determining working matrix, Filtering
- Identification and treatment of missing data
- Identification and treatment of outliers
- Identification and treatment of errors *(correct when possible)*
- Feature selection/extraction, dimensionality reduction
- Instance selection
- Data transformation
- Derivation of new variables

Feature selection

Evaluation of relevant variables in a dataset

- Priorization and ranking under different criteria
  - Feature weighting (determine weights of variables in the analysis)
- Elimination of irrelevant variables
  - Feature selection

Feature selection

- IA methods
- Statistical Feature selection: use statistical test for ranking
- Sometimes just use threshold on feature weighting ranks

Feature selection

- Goal: discard non-interesting variables
- Reduce data dimensionality
- Eliminate noise and redundancies
- Improve performance of algorithms
- Avoid spurious relationships in models
- Reduce curse of dimensionality
- Requires a response variable to be explained Y

- Rank relevance degree of Y wrt all other variables
- Discard less relevant

Statistical Feature selection

Guyon, I. (2008). Practical feature selection: from correlation to causality. NATO science for peace and security, 19, 27-43.

Hypothesis test:

H_0 : There is no relation between the y and x

H_1 : There is a relation

Get p-values for the dependence between Y and X

Lower p-values imply strongest dependence

Rank variables by ascending p-values

Discard irrelevant variables (threshold over p-values)

Specific tests depends on type of variables analyzed

Statistical Feature selection

Hypothesis test:

Y numerical

- X numerical: Correlations test / Sheffer generalized coefficient
- X qualitative: F test /Kruskal-Wallis

Y qualitative

- X numerical: F test/Kruskal-Wallis
- X qualitative: chi-2 test

Feature selection

Evaluation of relevant variables in a dataset

- Priorization and ranking under different criteria
  - Feature weighting (determine weights of variables in the analysis)
- Elimination of irrelevant variables
  - Feature selection

Feature selection

- IA methods (based on information theory)
- Statistical methods (based on statistical tests)
- Sometimes just use threshold on feature weighting ranks

AI Feature Selection

- Wrappers:

Rank subsets of features by accuracy in predicting Y (costly. Method specific oriented)

- Filters:

Rank subsets of features by some proxy measure (mutual information, statistical signifficance, Relief method)

- Embedded methods :

Implicit feature selection as part of the modelling algorithm, that penalizes less efficient variables internally (LASSO)

Relief

Score features

- For each feature X
  - For each object i
  - Find nearest neighbour from class A (j) and B (I), upon feature X
  - The score of the feature increases proportional to the distance of the neighbour in the same class of i and decreases proportional to the distance of the neighbour in the other class

    wk=wk+ d(i,j)-d(i,I) or wk=wk+d(i,I)+d(i,j) depending on class of I

  - Average final score modified by all objects

- Sort all X according to final scoring

Kira, Kenji and Rendell, Larry (1992). The Feature Selection Problem: Traditional Methods and a New Algorithm. AAAI-92 Proceedings.

Preprocessing

Data cleaning
Data preparation
Data preprocessing

- Formatting issues, building software context
- Determining working matrix, Filtering
- Identification and treatment of missing data
- Identification and treatment of outliers
- Identification and treatment of errors *(correct when possible)*
- Feature selection/extraction, dimensionality reduction
- Instance selection
- Data transformation
- Derivation of new variables

Variables Transformation

- Homogeneization
- Approaching to methods hypothesis
- Getting more interpretability

Variables Transformation

- Homogeneization
- Approaching to methods hypothesis
- Getting more interpretability

Variables Transformation

- Data cleaning reasons
  - Measurement units of Thyroids hormones from different laboratories

  1993
  Collaboration UPC, Barcelona, Spain
  Andrija Stampar School of Public Health, Zagreb, Croatia
  Setre Milordsnice Clinical Hospital, Zagreb, Croatia

  Find patterns of thyroids dysfunctions 1002 patients, 12 measurements

  2013
  Collaboration UPC, Atención Primaria ICS
  http://www.sidiap.org/

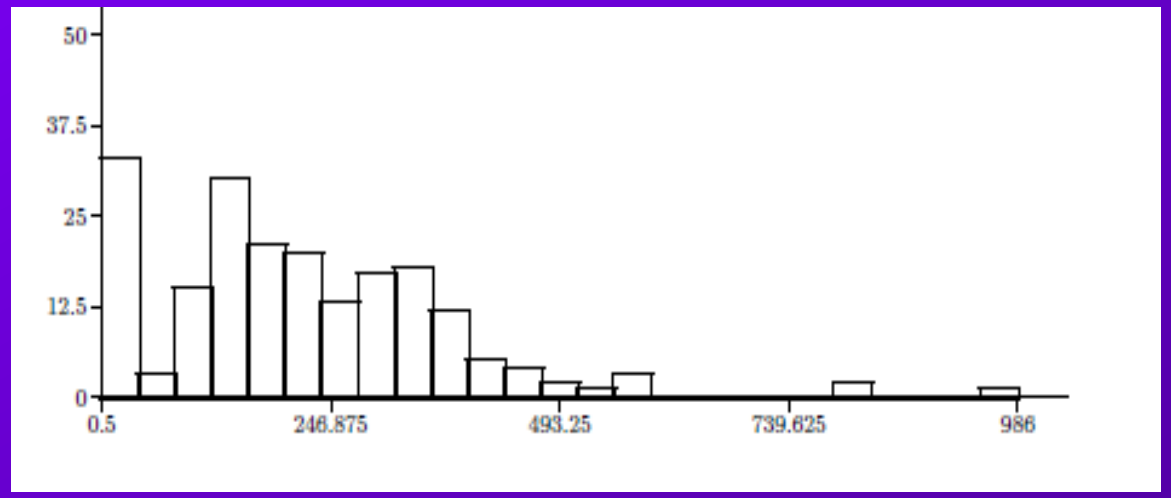
  Laboratory measuments in TSH

- Spatial incident due to mistake in measurement units
  https://www.youtube.com/watch?v=kxWui3BEgGc

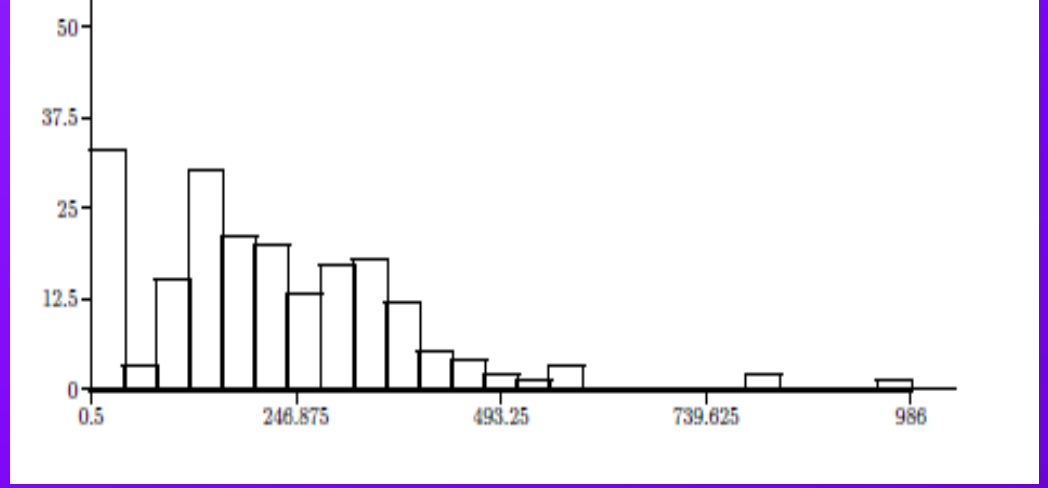
Laboratory Tests measurements

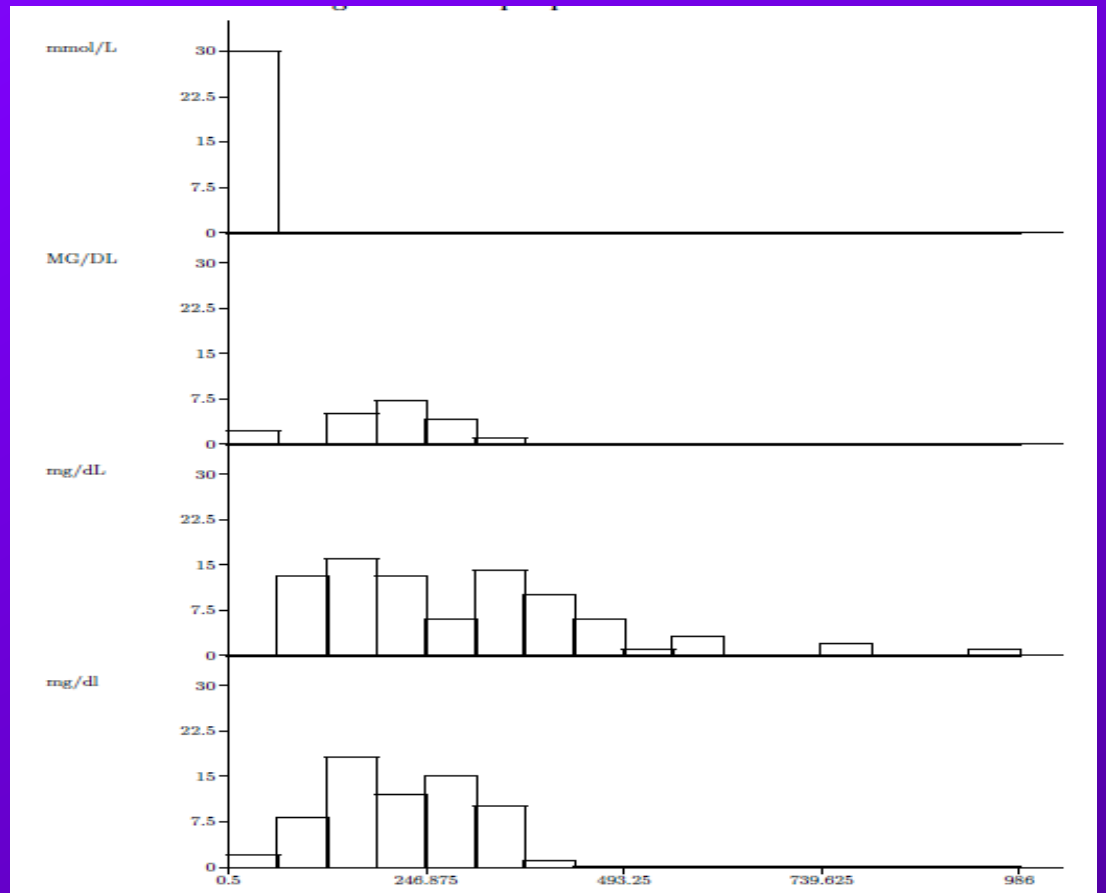
Measurement of Total Cholesterol from 200 pacs from Catalan Public Health System in 2013 (Primary Care)

Summary Statistics	
Number of objecs	200
Number of missing values	0
Number of useful values	200
Mean	213.2037
Median	193.9
First Quartile (Q1)	115.8
Third Quartile (Q3)	306.2
Minimum	0.5
Maximum	986
Quasi-standard deviation	156.4218
Variation Coeficient	0.7318



Frequency Table		
Modalities	Freq. absol.	Freq. relat.
mg/dl	66	0.33
mg/dL	85	0.425
MG/DL	19	0.095
mmol/L	30	0.15
missing data	0	0





Unitat	Count
(mg/dL)	12
0.0 - 240.0	1
55519	1
G/L	4
mg/dl	50
mg/dL	50
MG/DL	50
mg/dl^MMOL/L.	50
mmol/L	50
Mmol/L	1
MMOL/l	1
MMOL/L	1
mmol/L^mg/dL	50
N=	321

mmol/l = 38,669 mg/dl

Variables Transformation

- Data cleaning reasons
  - Measurement units of Thyroids hormones from different laboratories
  - Refer the whole set of variables to comparable units
    all concentration variables in mg/l
    proportions instead of absolute numbers, ...
  - Coertions: Information loss.
    - Discretization (h/week working)
    - Categorization (Thiroids levels)
    - Recategorizations (professions)

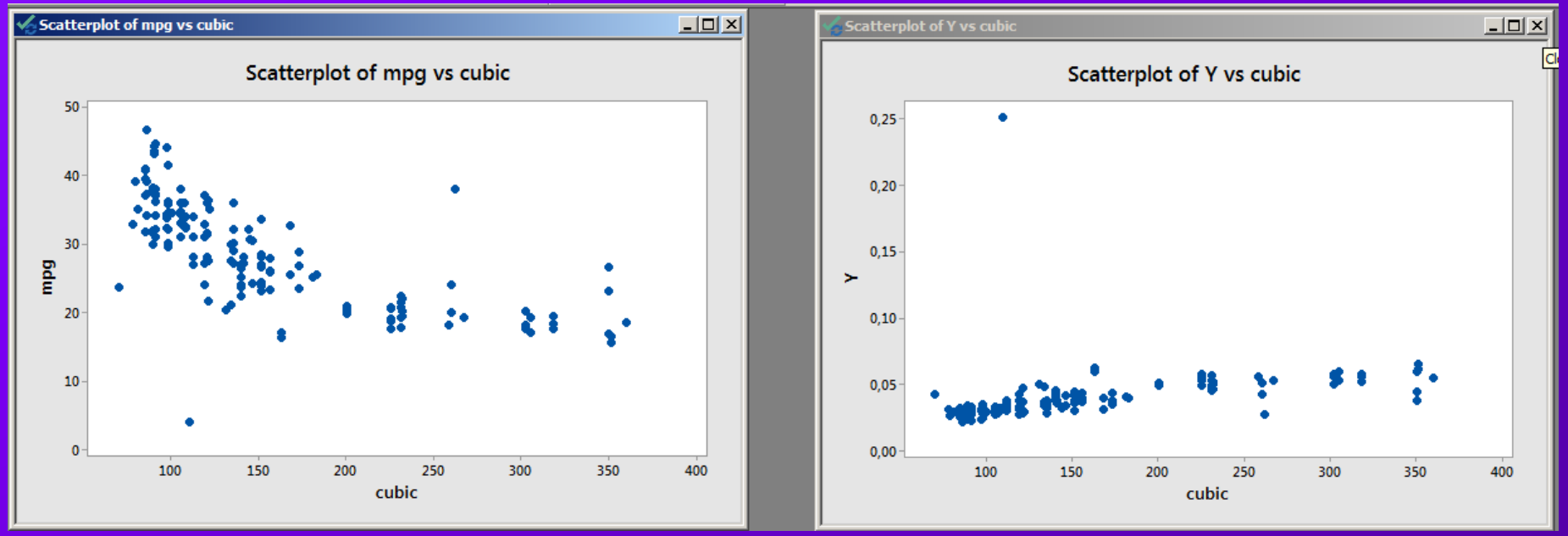
Better avoid

- Technical questions:
  - Estandarditzation, normalitzation o linealirization
  - Eventual logaritmic transformation
  - Required by data mining technique to apply

Select a technique respectfull with original data

Exceptional situations

where transforms make sense



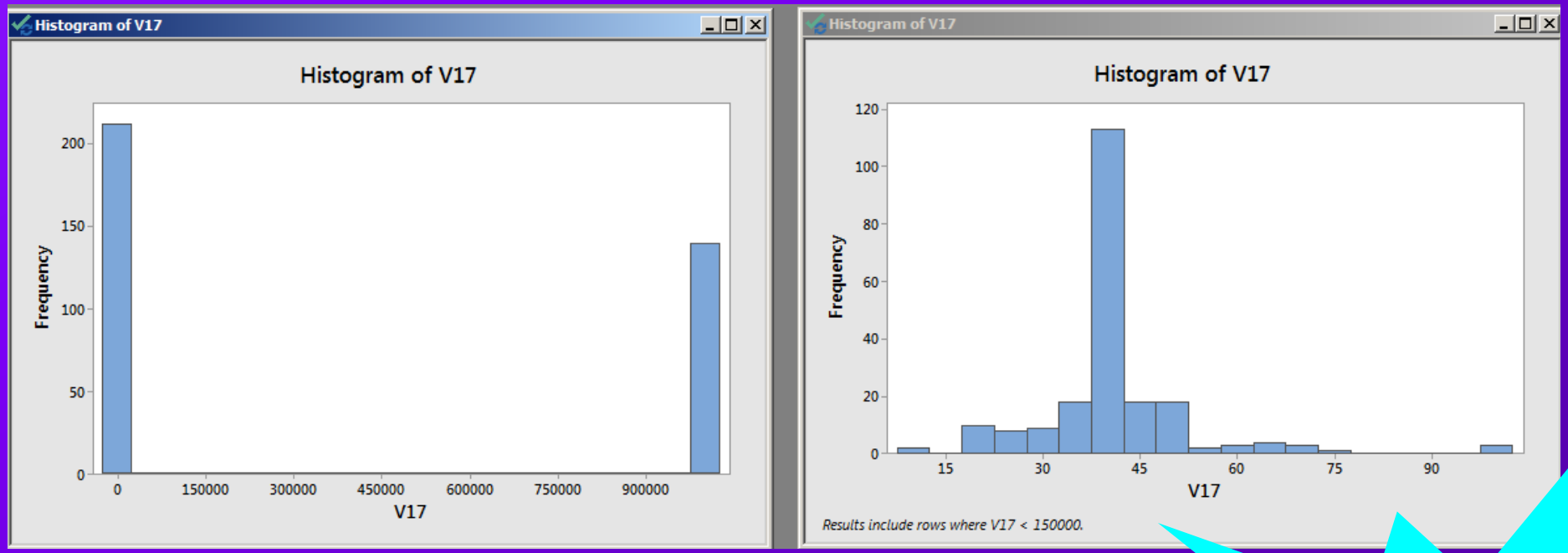
- Mpg: miles per gallon of a car
- Cubic: cubic capacity of the car engine
  Non linear relationship (regression non suitable)

- Y‘ = 1/Y : Linearizes the relationship

Y is car Consumption!!!!

Exceptional situations

where transforms make sense



- Hours working per week
- 3-modal:
  - Arround 20 h/w
  - Arround 40 h/w
  - Arround 65 h/w

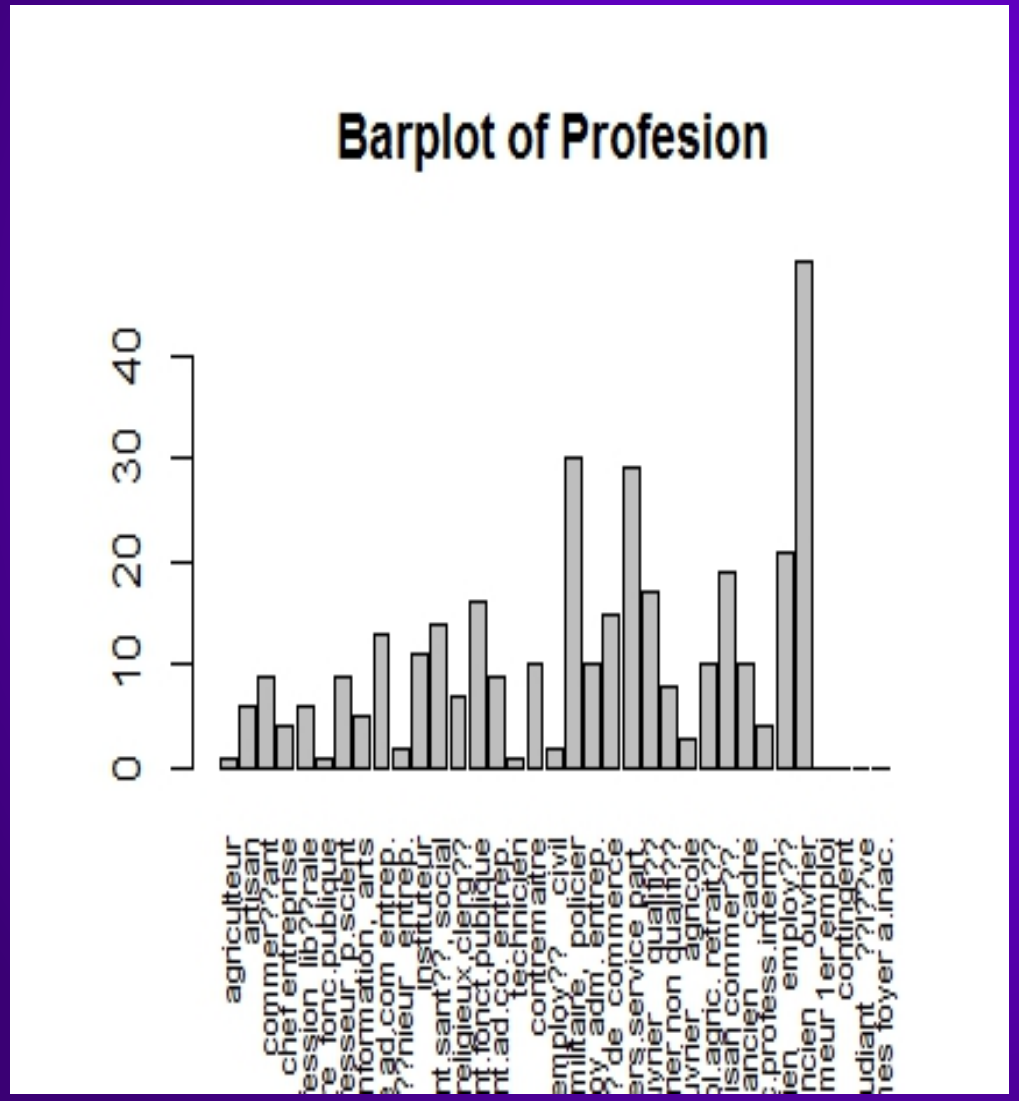
- Correspondence with part-time, full-time, extra turns works

Build a qualitative variable: Type of work (part-time, full, turn)

Exceptional situations

where transforms make sense

Regroup professions in a more general families



- Professions: 31 modalities unmanageable
- Families of professions:
  - agriculteur
  - ouvrier agricole
  - expl.agric. Retraitée

  Agriculture sector
  - Artisan
  - anc.artisan commerce
  - information, arts
  - commerçant
  - employée de commerce

  Arts and commerce

Preprocessing

Data cleaning
Data preparation
Data preprocessing

- Formatting issues, building software context
- Determining working matrix, Filtering
- Identification and treatment of missing data
- Identification and treatment of outliers
- Identification and treatment of errors *(correct when possible)*
- Feature selection/extraction, dimensionality reduction
- Instance selection
- Data transformation
- Derivation of new variables

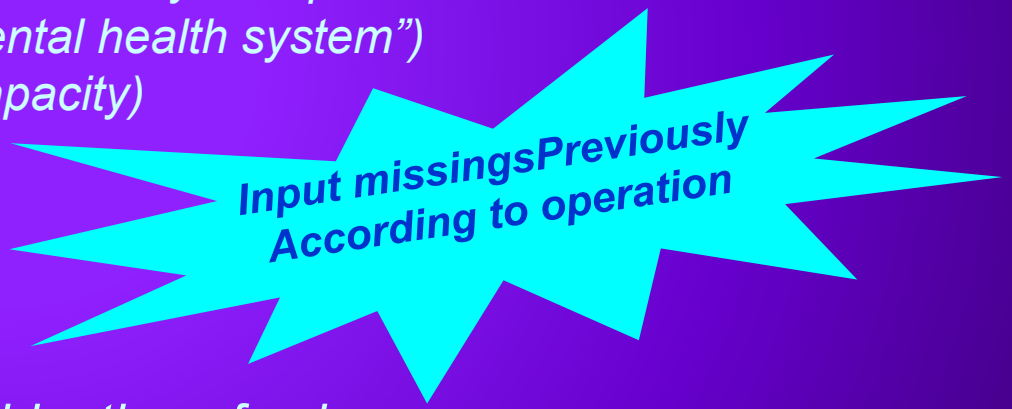
Derivation of new variables

- Aggregates (additions of other variables)
  - Total household income

- Synthetic indicators
  - Classical generation of global score in psychometric scales
  - Indicators

  (Lund parameter =external contacts/days hospital indicator of "approach of a mental health system")

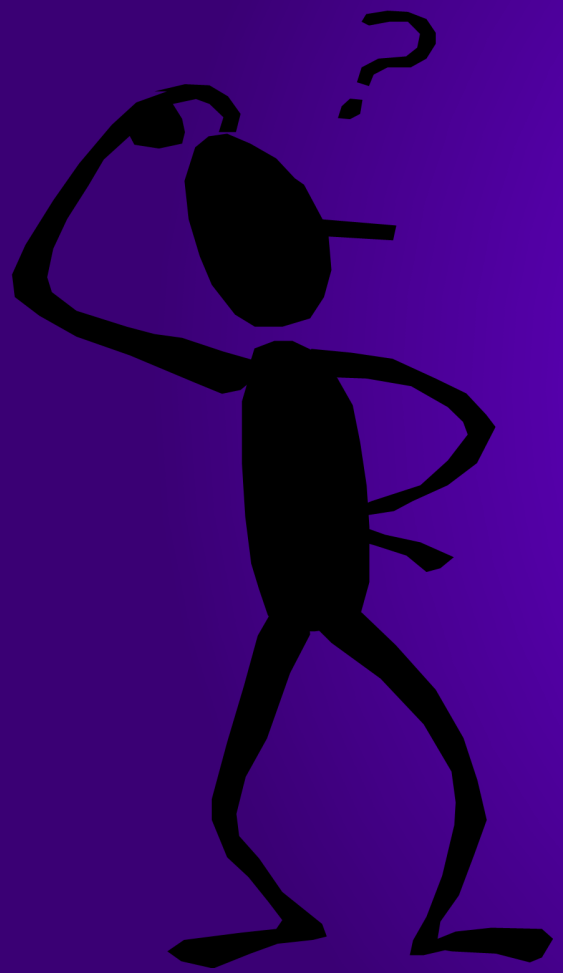
  Case Credit Scoring (saving capacity)



- Binary indicators
  - *If condition regarding a combination of values then indicatior=1, else the indicator=0*

  (fastDummies Package, dummy_cols()) (Package dummies dumy.data.frame())

- *Dimensionality reduction techniques*



Any question?...

K. Gibert

https://www.eio.upc.edu/en/homepages/karina

karina.gibert@upc.edu @Karinagibertk

KEMLG-@-IDEAI: Knowledge Engineering and Machine Learning group at Intelligent Data Science and Artificial Intelligence Research Center

Universitat Politècnica de Catalunya, Barcelona

Dean of Illustrious Professional College of Informatics Engineering of Cataloonia (COEINF)

Founder and president of donesCOEINF