Preprocessing

K. Gibert

Knowledge Engineering and Machine Learning group at
Intelligent Data Science and Artificial Intelligence Research Center
&
University Research Institute on Science and Technology for Sustainability
&
Dep. S*tatistics and Operations Research*

Universitat Politècnica de Catalunya-BarcelonaTech

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

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
Gibert, K (2009) Estadística: Contexto histórico e introducción a la descriptiva 4a. ed. Serveis Gràfics Copisteria Imatge S. L. Feb 2009. DL: B-10513-2009.
Gibert K (2003) Introducción a la Estadística Descriptiva. Ahlens S. L. May 2003. DL: -27564-2003

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)*

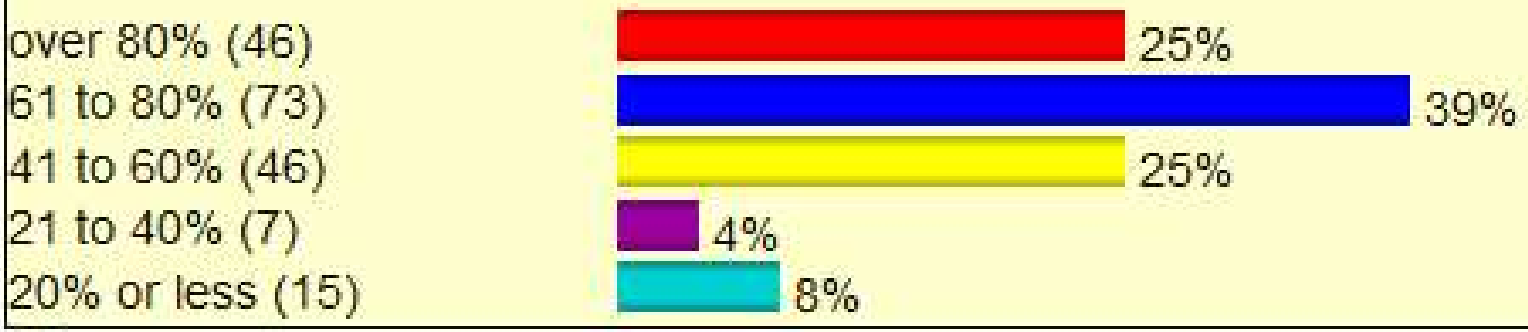
Impact of Preprocessing in real Data Mining projects

Data preparation part in data mining projects

Share Tweet

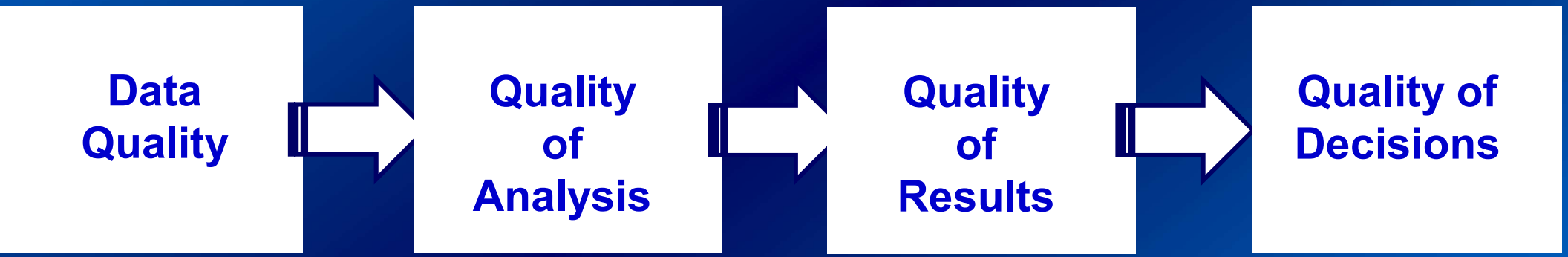
Poll

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



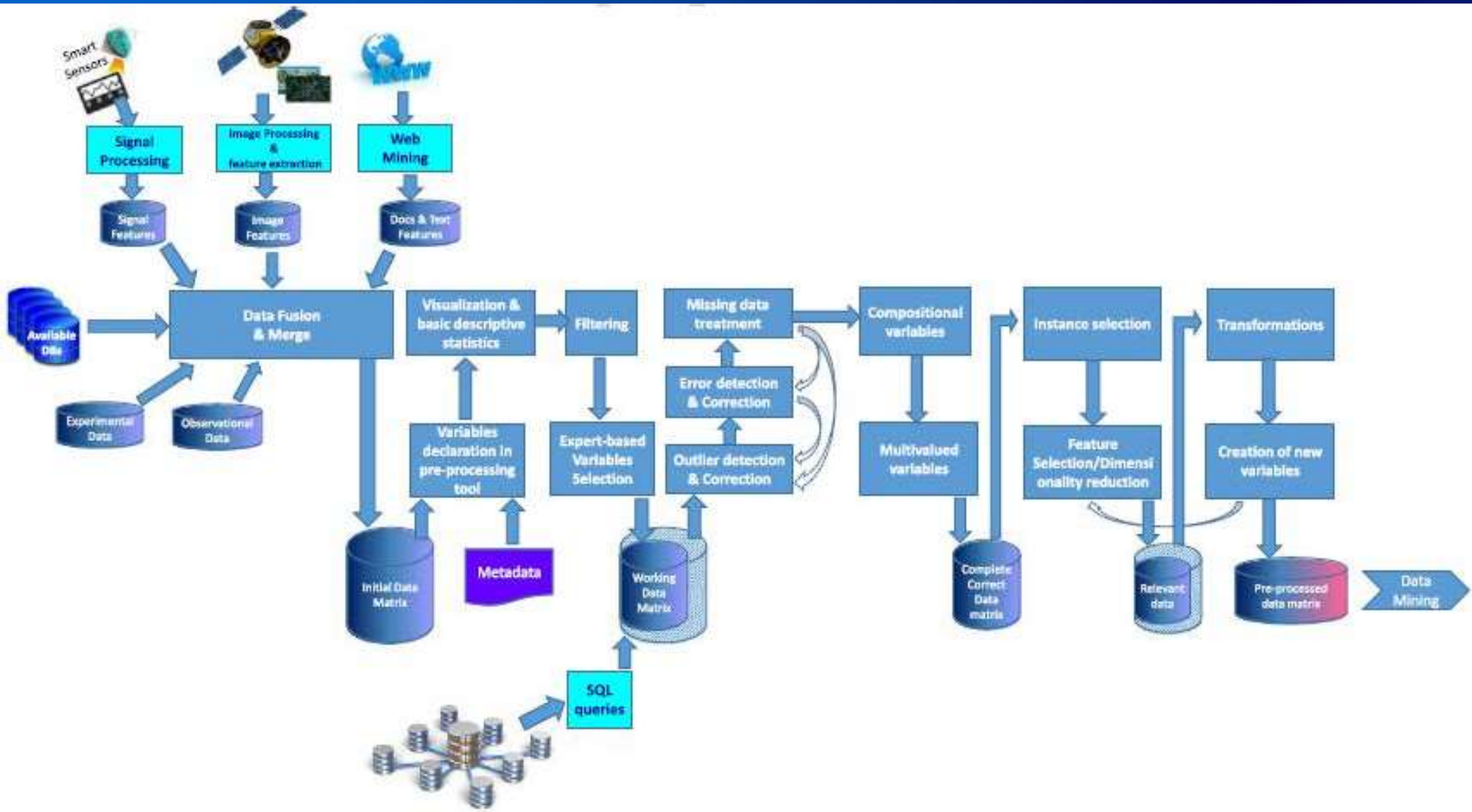
http://www.kdnuggets.com/polls/2003/data_preparation.htm

Preprocessing



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

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

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*
$\hat{\sum}_j$ *maximum likelihood estimate of the covariance matrix*
rj number of complete variables for pattern j
K total number of variables

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

The Little test in R

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

  x: dataframe, matrix less than 50 variables

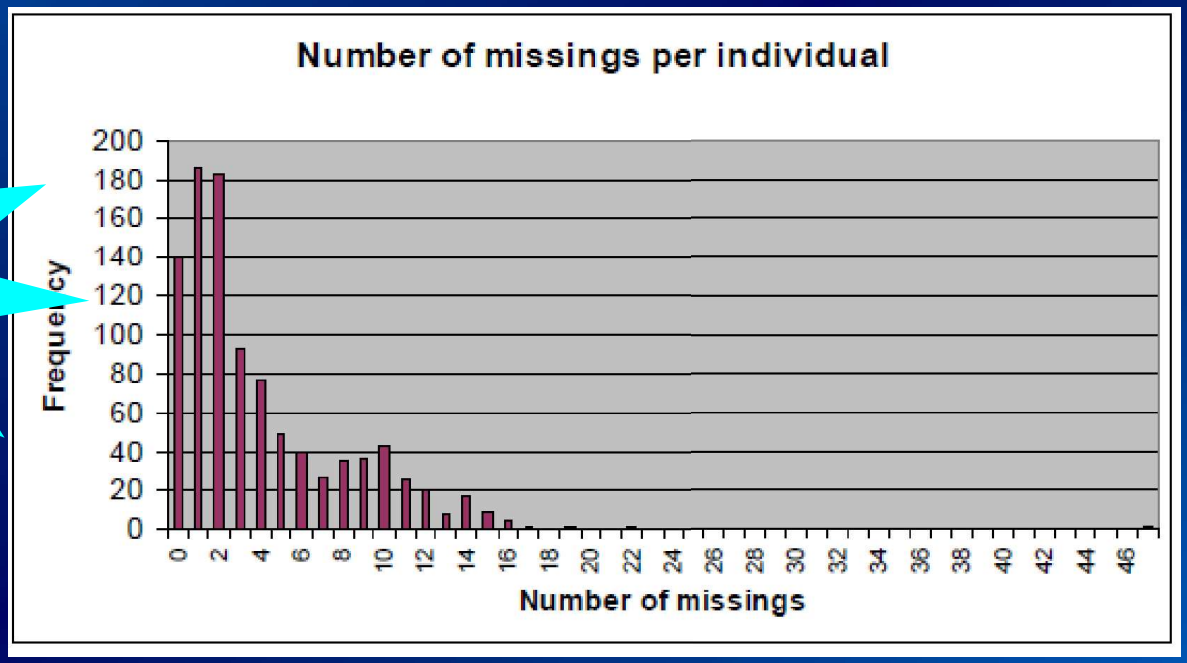
- **Returns:**

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



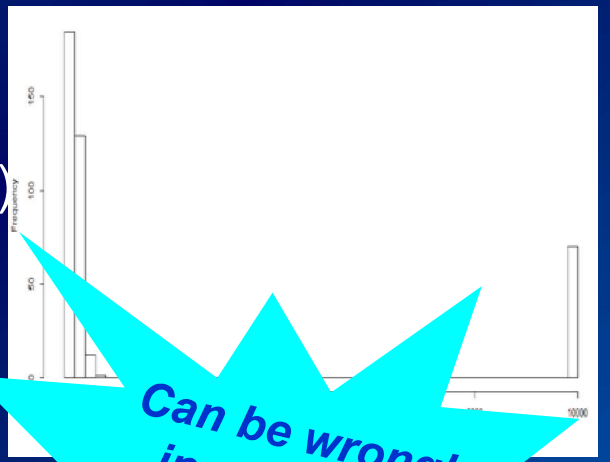
Reliability of individual

- 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 roc
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

SPLIT

Cures Full

Cures Missing

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 roc
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

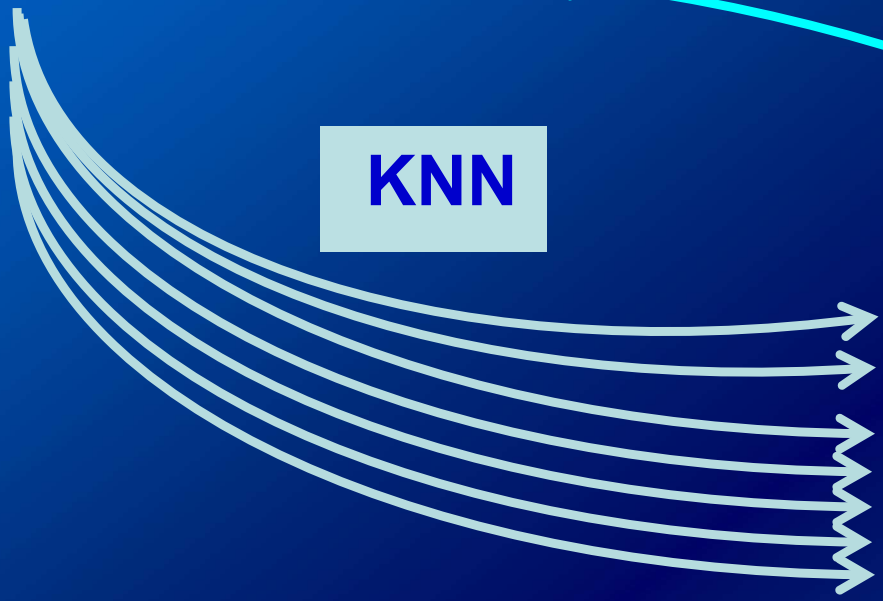
C_HISTORI	C_TRACTAI	DATA	Alimentació	Cures d'aparenca	Higiene	Vestit: part superior	Vestit: part inferior	Utilització del bany	Bufeta	Intestí	Llit, cadira, cadira de rod
1645,0	84990,0	21/07/2003 0:00	7			6	2	6	6	6	3
1858,0	76281,0	26/09/2002 0:00	7			7	7	6	5	5	7
2267,0	86796,0	01/10/2003 0:00	7			7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

KNN

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 roc
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	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	7	6	6	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 roc
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*

Inputation:

Complete the 42x16 data matrix

Clustering the full matrix

Hierarchical

Ward criterion

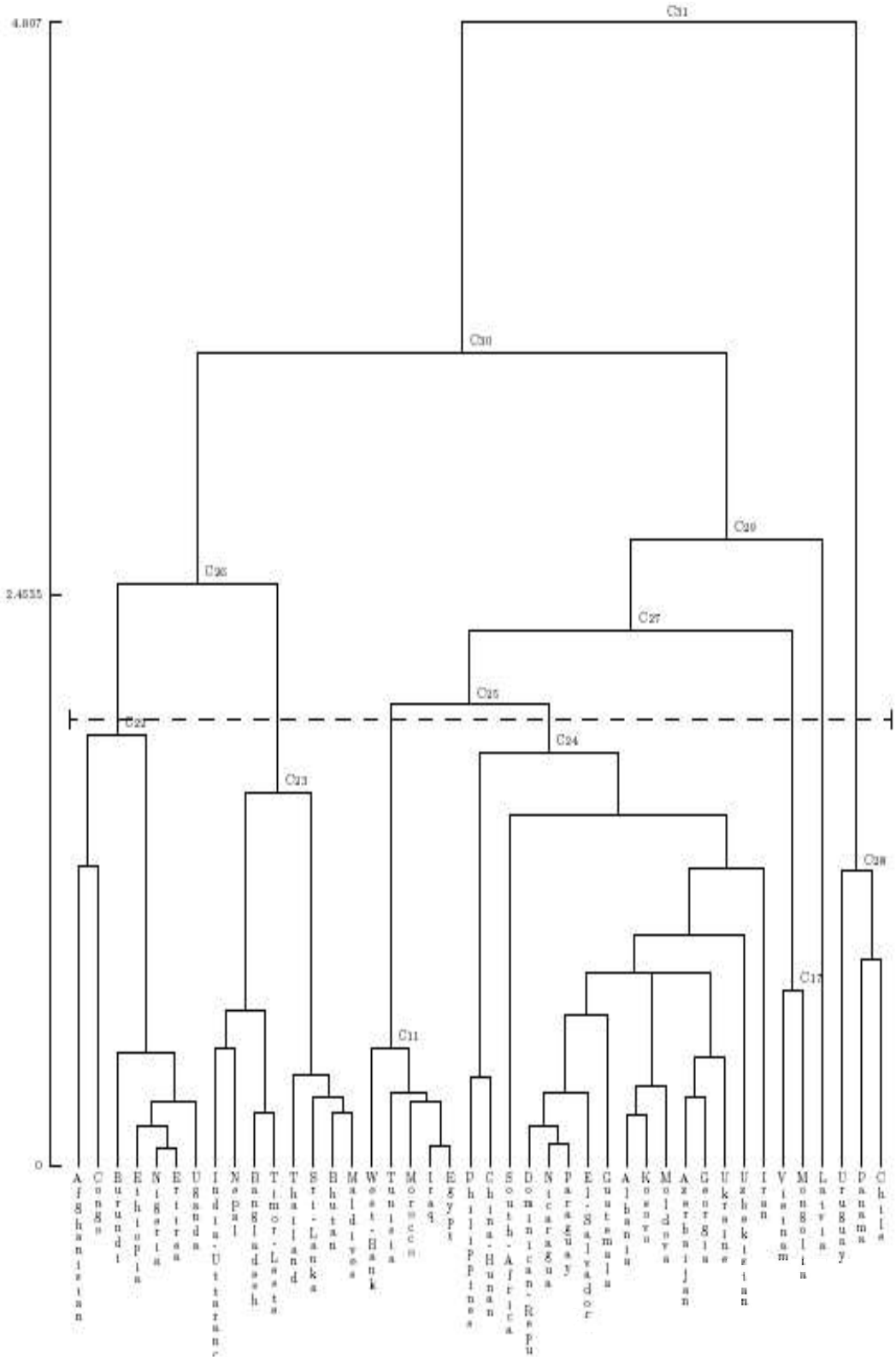
Gibert's mixed metrics

[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]*

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
comcarewor	$\bar{X}$	0.0314	0.0856	0.0197	0.0269	0.624	0.1991	0.1313
	S	0.0083	0.0196	?	0.0067			
	N*	5	10	7	1	4	0	1
usmhexperca	$\bar{X}$	0.2646	0.4951	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

MICE method

[vanBuuren1999]

multiple imputation by chained equations

Multiple imputation (MI):

- Replace missing values with *plausible* substitutes
  - Distribution-based
    maximum-likelihood based Markov-chain Monte Carlo (MCMC)
  - Inject *the right amount of randomness* to reflect uncertainty
- Repeat *m* > 1 times to procude m imputed datasets
- Analyse datasets individually, but identically
- Combine the models, get confidence intervals using Rubin's rules (`micombine`)

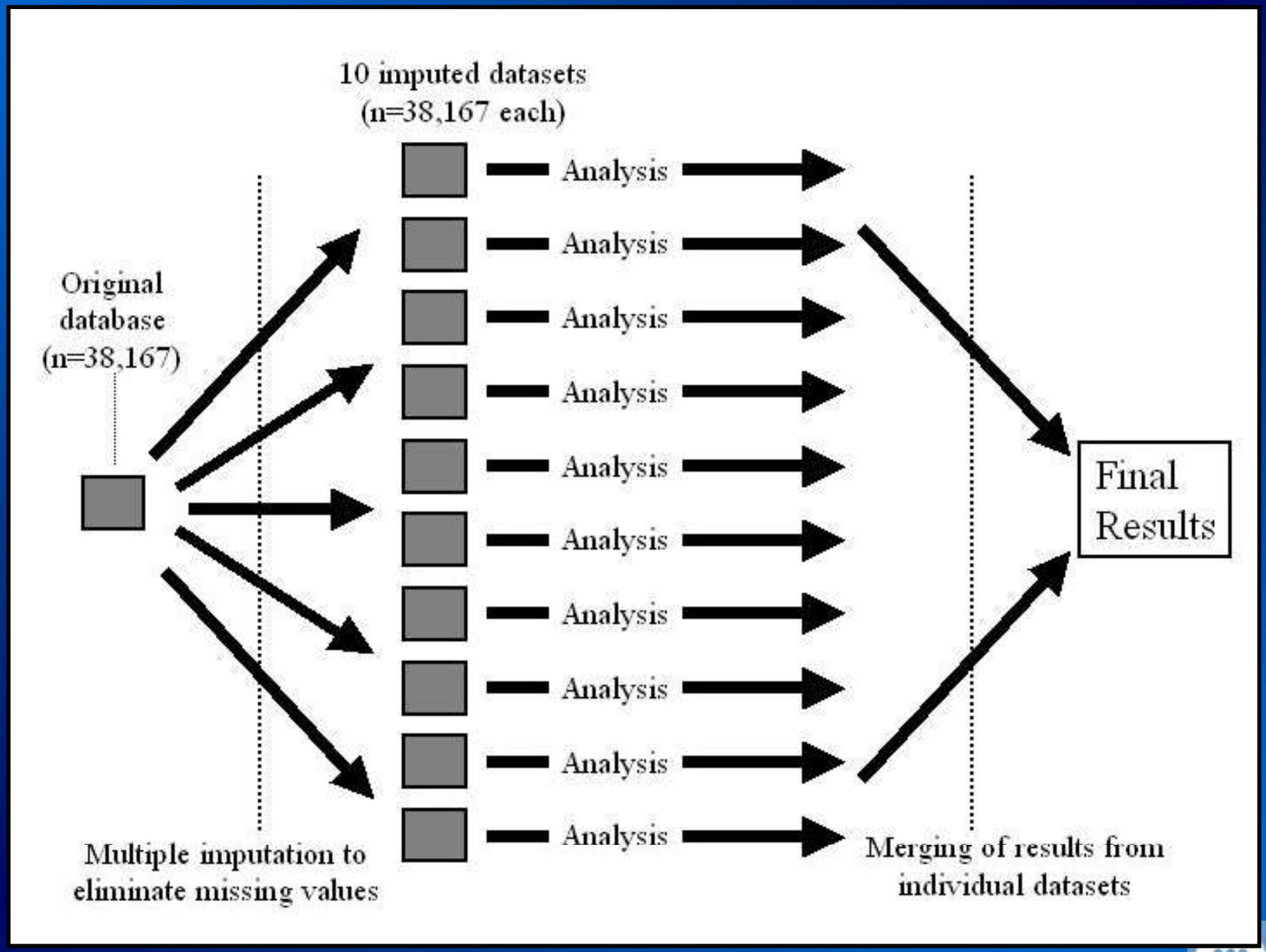
Context: Multiple regression In general

The MICE approach has three components:

- Univariate – implemented in `uvis`
- Multivariate – implemented in `ice`
- Multiple – implemented in `ice`

- `ice` = imputation by chained equations

MICE



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.

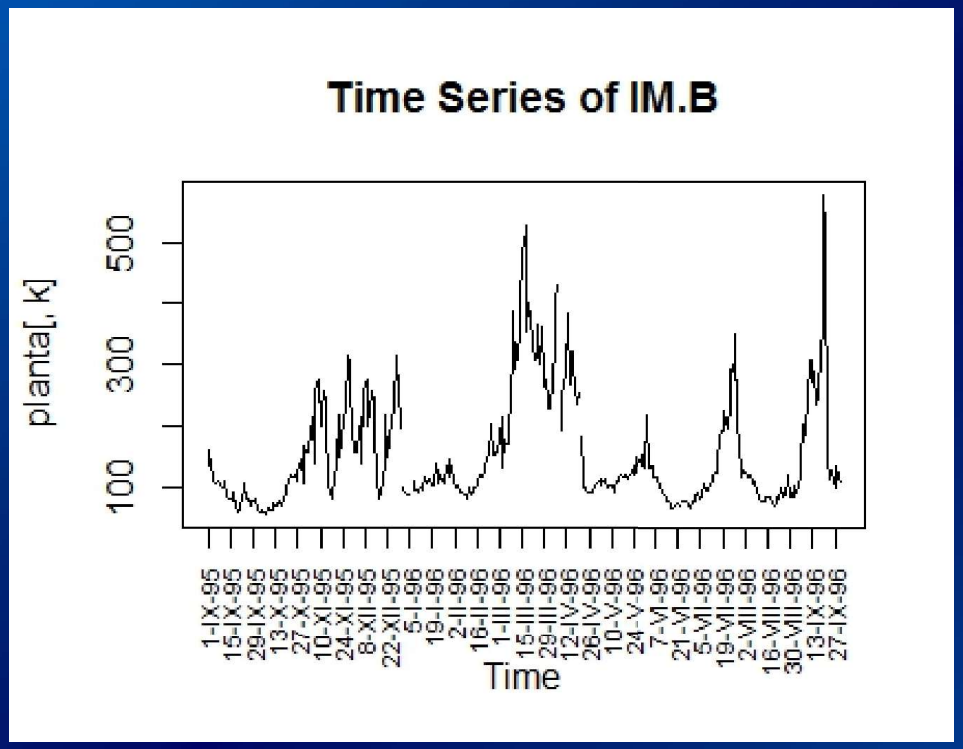
Interpolation

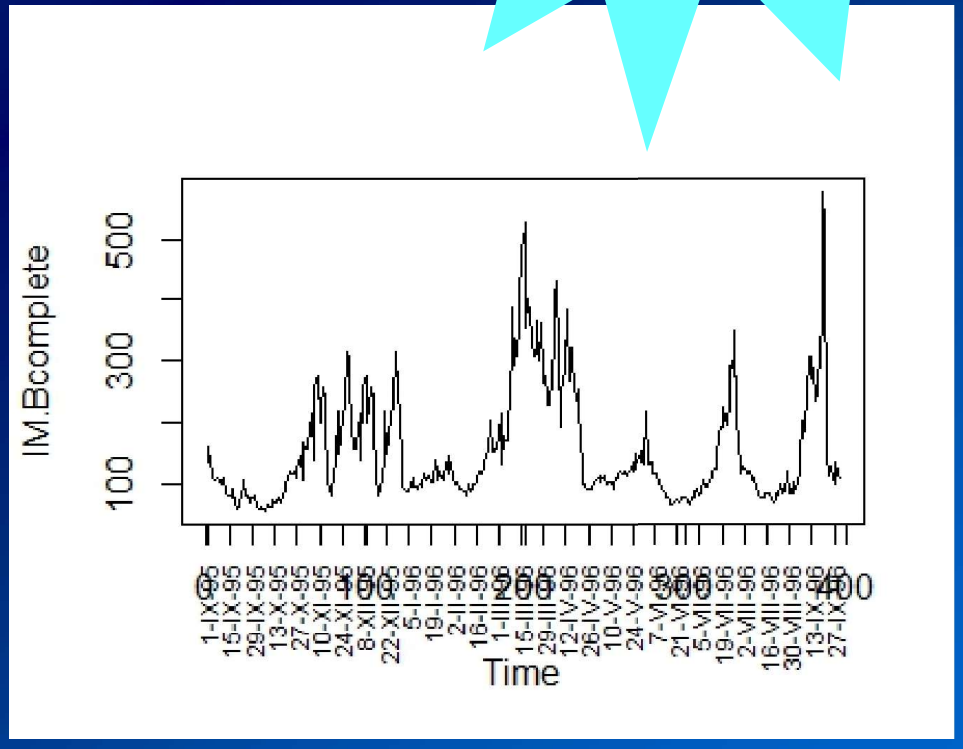
Usefull for time-series of numerical va

- Linear assumption between observed points

(assume monotonic behaviour between observa

ALTERNATIVE
Assume constant between Measurements (slow dynamics)





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

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, weighing
  - *Treat apart. CLEARLY REPORT ABOUT IT*
- Missing code
  - *Substitute by missing or inpute*

Care with suppressions

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

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)

Variables Transformation

- Homogeneization
- Approaching to methods hypothesis
- Getting more interpretability

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)

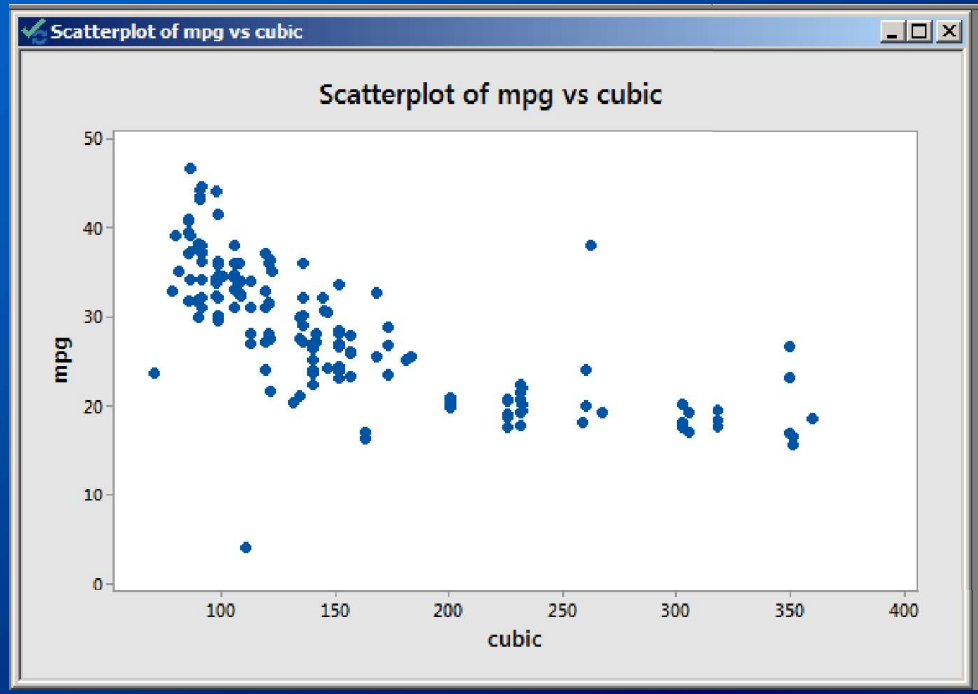
Better avoid

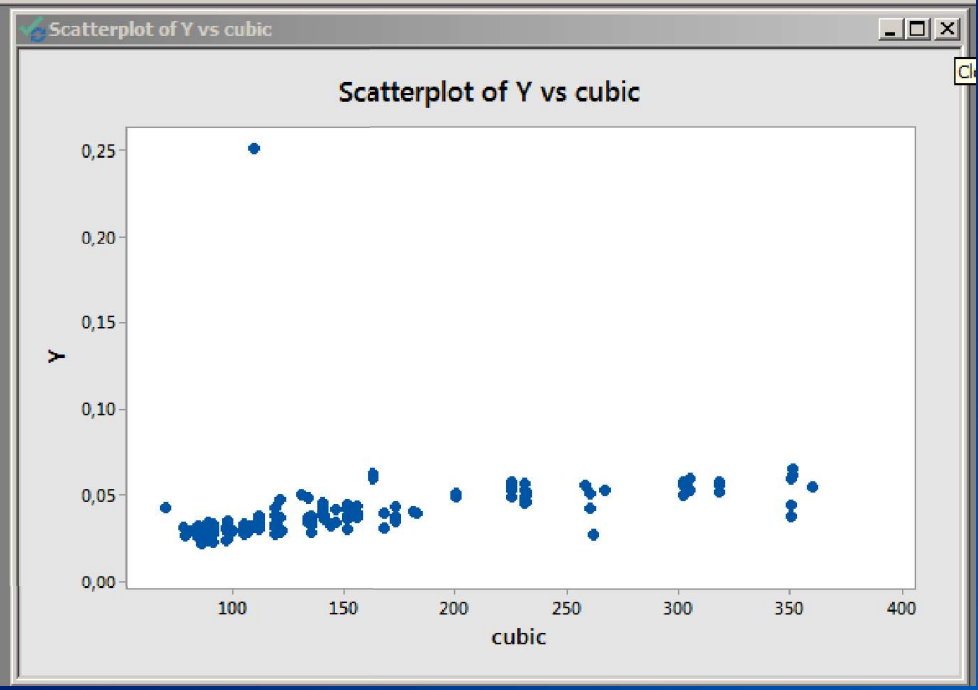
- Technical questions:
  - Estandarditzation, normalitzation o linealirization
  - Eventual logaritmic transformation
  - Required by data mining technique to apply

Select a technique respectfull with original data

Exceptional situations

where transforms make sense





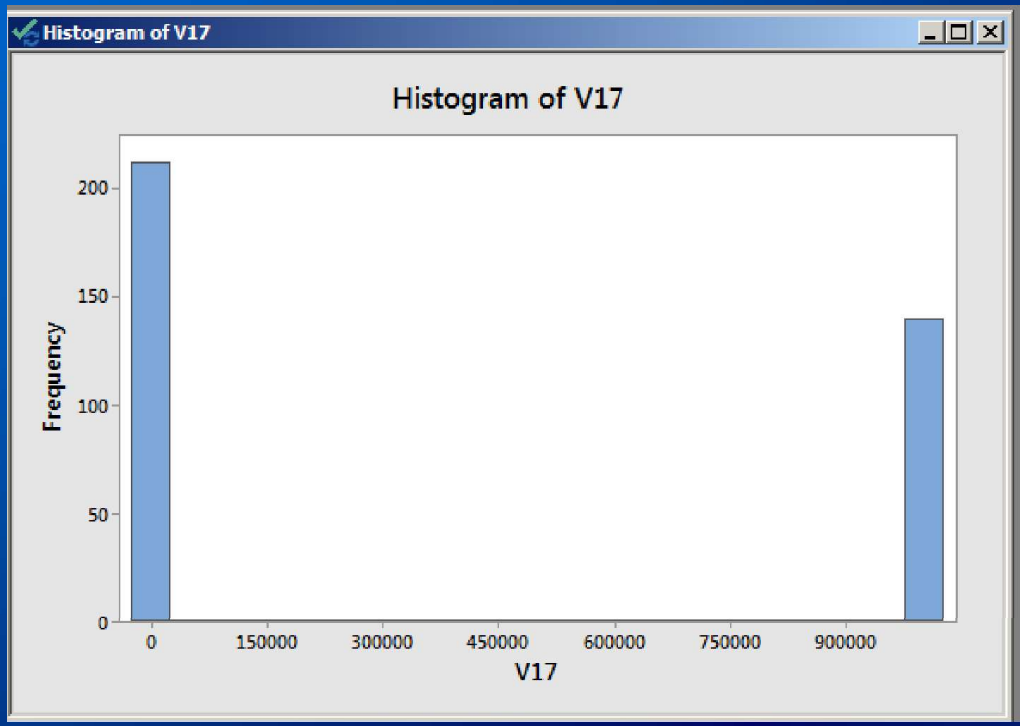
- Mpg: miles per gallon of a car
- Cubic: cubic capacity of the car engine
  Non linear relationship (regression non suitable)

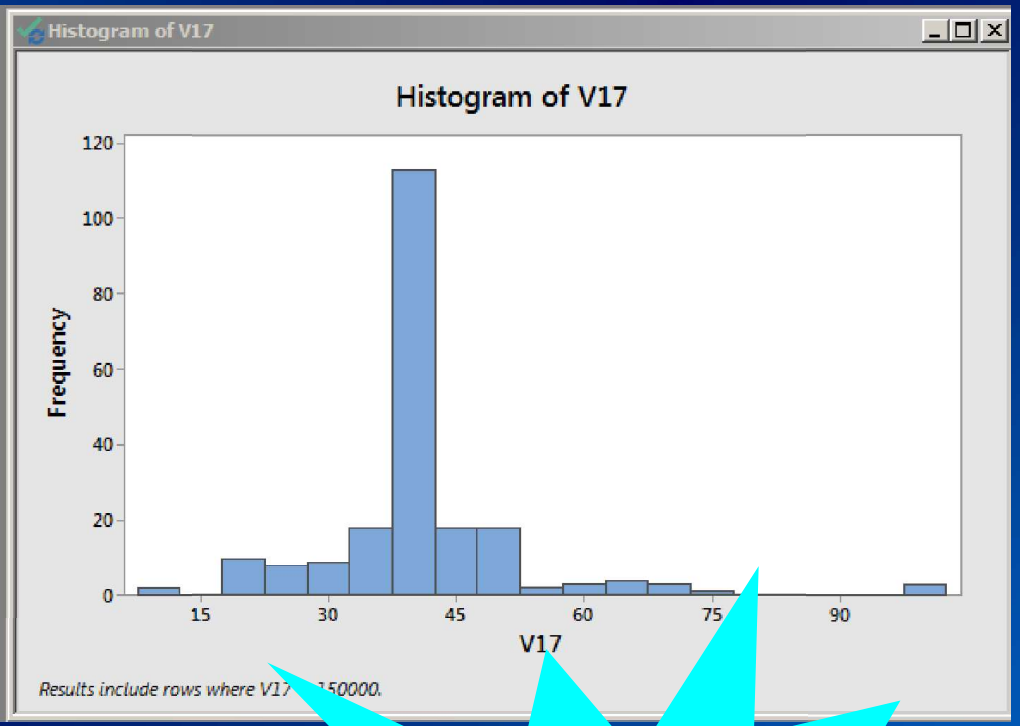
- Y = 1/mgp : Linearizes the relationship

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

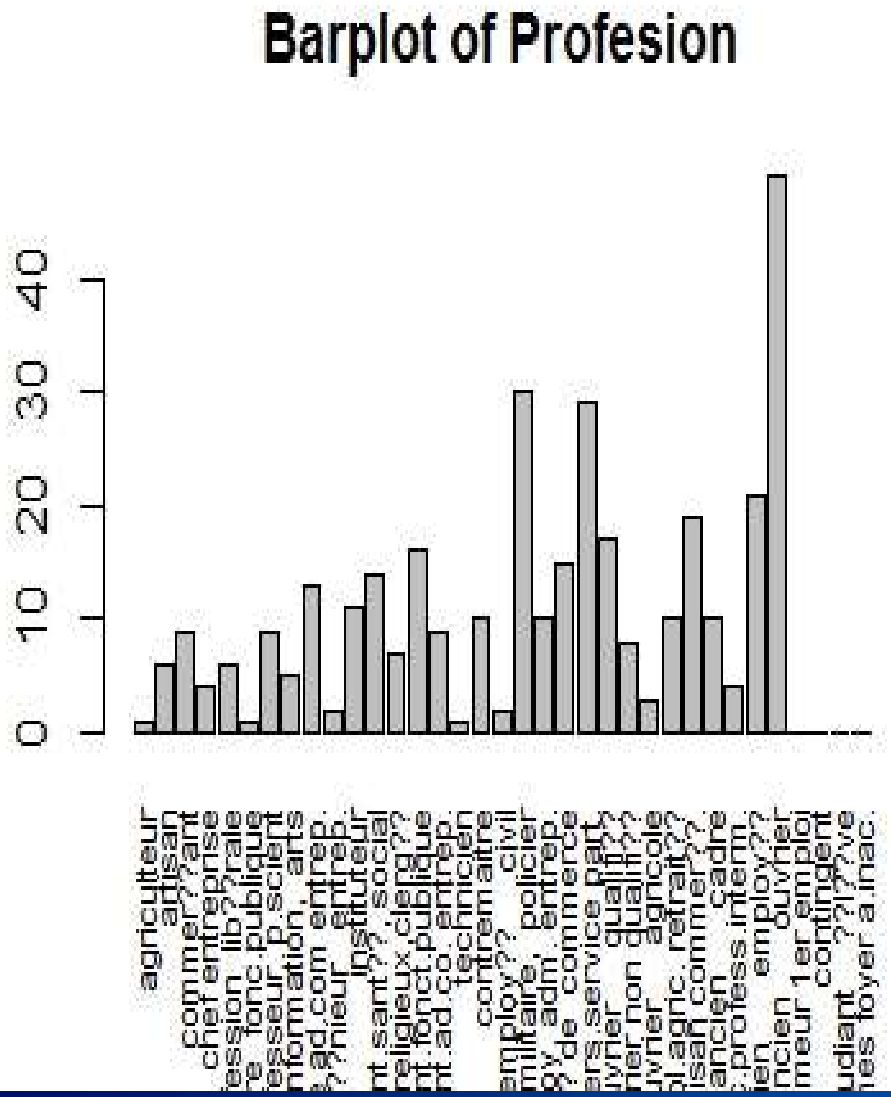
- Correspondence with part-time, full-time, extra turns works

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

Exceptional situations

where transforms make sen

Regroup professions in a more general families



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

  Agriculture sector

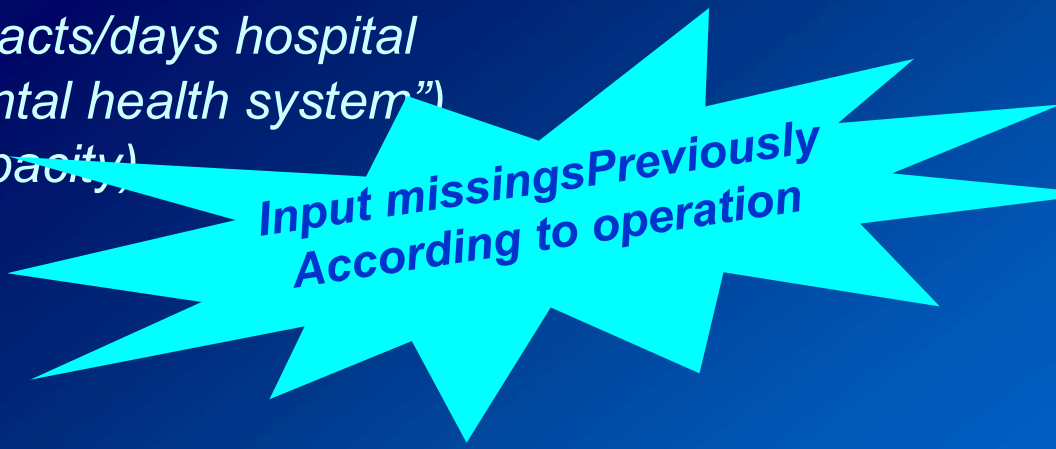
  - Artisan
  - anc.artisan commerce
  - information, arts
  - commerçant
  - employée de commerce

  Arts and commerce

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*
- *Dimensionality reduction techniques*

Datos, Descriptiva y Pre-processing

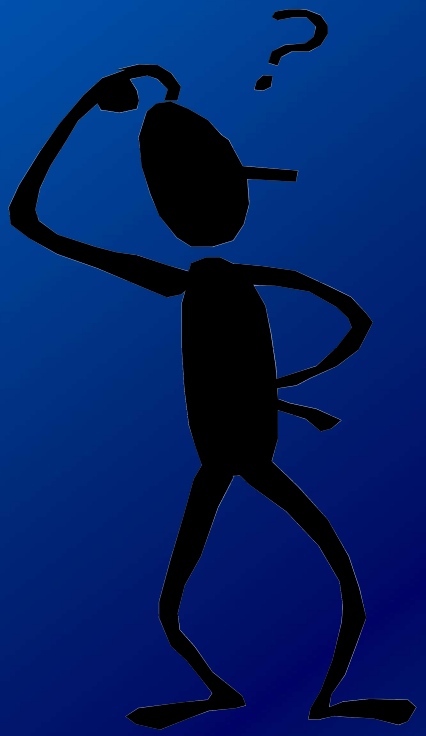
Karina Gibert

Dpt. Statistics and Operation Research

Knowledge Engineering and Machine Learning Research group at
Intelligent Data Science and Artificial Intelligence Specific Research Center

Institut Universitari de Recerca en Ciència y Tecnologia de la Sostenibilitat
Universitat Politècnica de Catalunya-BarcelonaTech (Spain)

karina.gibert@upc.edu
www.eio.upc.edu/homepages/karina



Are there any questions?...