Grado de Inteligencia Artificial - GIA

Advanced Preprocessing – Missing Data & Feature Engineering

Prof. Dante Conti

Prof. Sergi Ramírez

Credits: Tomàs Aluja-Banet

Missing data

- **Databases:**
  - Databases are used to extract knowledge, but only information that is currently used is maintained. → Enough?
  - Not compulsory fields.
  - Errors and outliers may be taken as missing values …

- **Surveys:**
  - Outright refusals: *unit nonresponse* →change the sample
  - Nonresponse to some items: *item nonresponse* → dealing with missings (it depends on the data collection method: internet, telephone, mail, face to face)
  - Inapplicable questions to some respondents → this is not missing data
  - Dropouts in panel studies → Censored data

Serious drawback of the data quality (values not recorded, not consistent, …)

Missingness is a nuisance

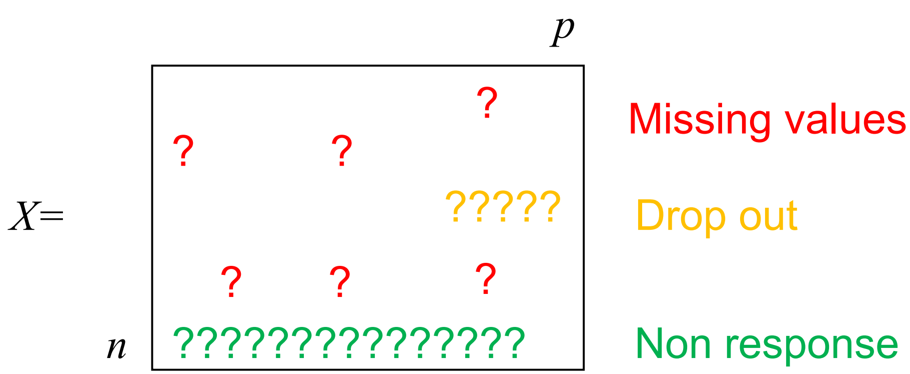
The missing data problem

- Number of missing values is an indicator of the data quality
- Missing data can appear in several forms:

  <empty field> "0" "." "99999" "NA" …
- Standardize missing value code(s)

- Typical data set:

Some information is missing for some variables and for some cases.



- Analysis is just designed for complete data sets (standard methods will fail)

Is missing data a problem

1. Ignoring missing data can seriously bias the results

2. Missing data represents a loss of information (waste of resources)

3. The impact of missing data depends on its generating mechanism (why some values are missing?)

The best policy to deal with missing data is to avoid it with careful planning of data collection, with proper intelligent interfaces.

Activity # 1 - Dealing with missing data

Before to start. Identify the missing data

Usual convention:

Assign a missing code to continuous variables (NA, -1, 999999, …)

Assign a new category (missing) to a categorical variable.

Check the quality of the information

Count the number of missings per variable and rank them accordingly.

The more the missing the less reliable is the information provided by the variable

Characterize the missingness mechanism

Create a new variable counting the number of missings per individual.

Profiling of global missingness analysis (individual) and compare both groups of cases (variables & individuals)

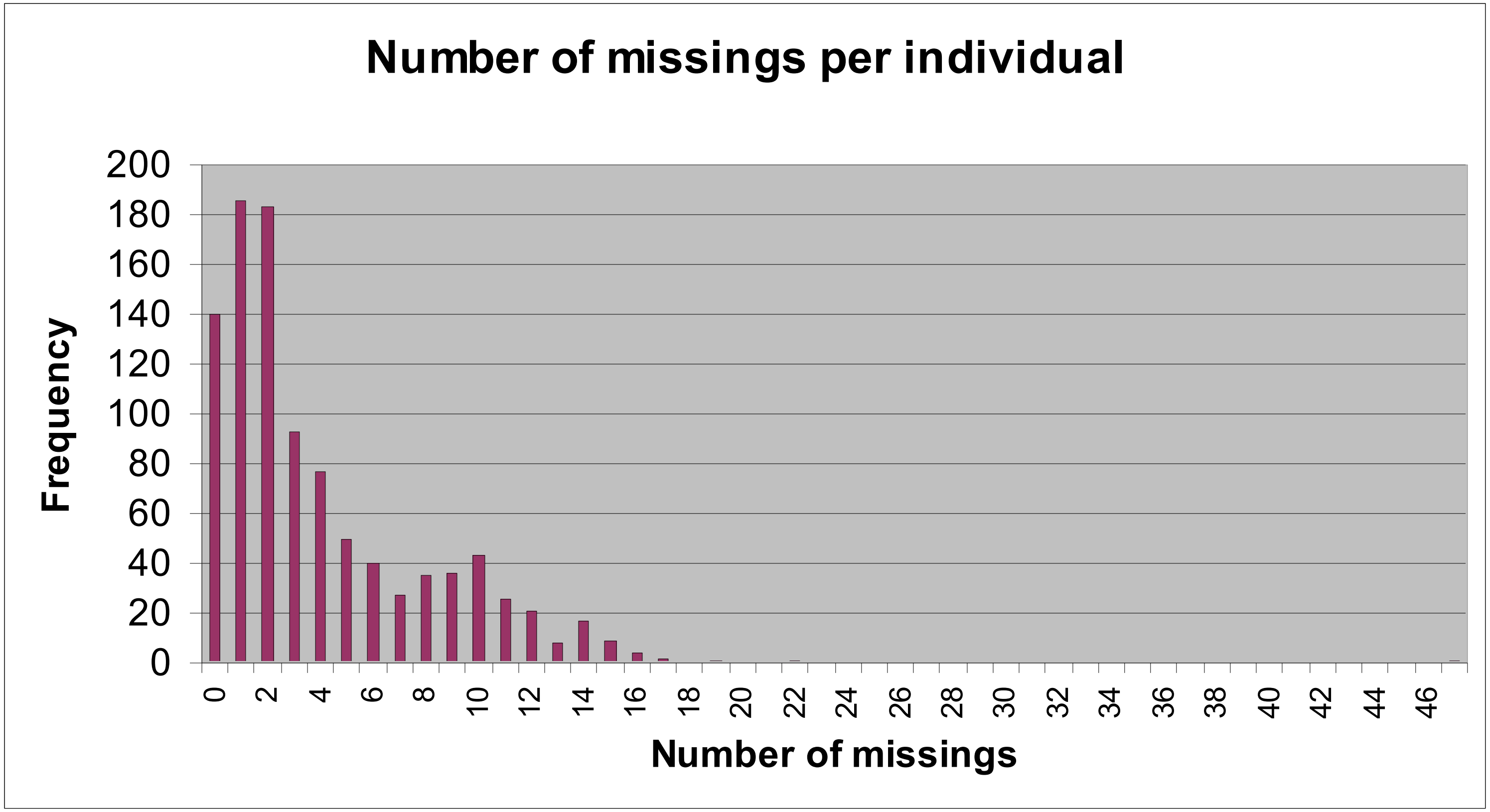
Dealing with missing values

Ignore records with missing values

If categorical, treat missing value as a separate level (or impute them).

If continuous, impute (fill in) with mean or median values, 1nn, …

Global analysis of missingness



Missingness mechanisms

- MCAR - Completely at random: missing values appear without any pattern. This is the most favorable situation.

- MAR - At random: missing values appear related to third observed variables. This is the most usual case, i.e., asking the income of individuals, income is missing but can be imputed from the educational level.

- MNAR - Not at random: missing values depend on the missing variable itself. This is the most difficult case. In the previous example it would be that high incomes tend to not declare it.

Missingness mechanisms

Missingness mechanisms

MCAR and MAR →Imputation

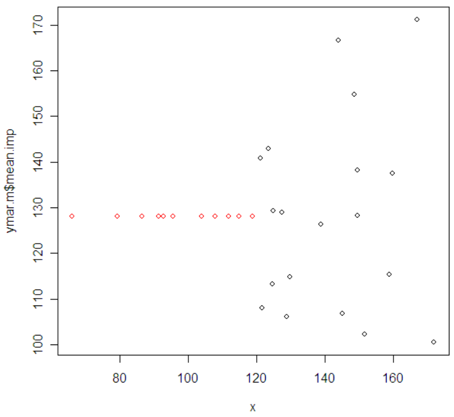
MNAR → Be careful (STOP)

IMPORTANT https://search.r-project.org/CRAN/refmans/naniar/html/mcar_test.html

Treatment of missing values

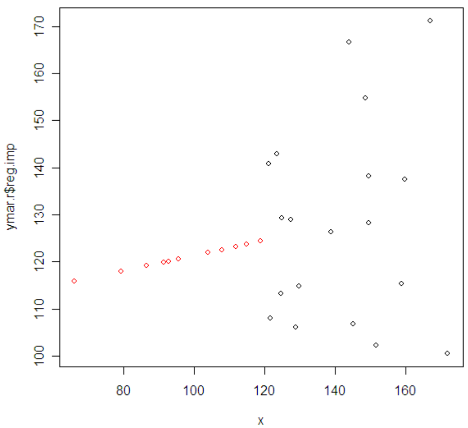
Traditional methods

- Listwise deletion. Every individual with a missing value is deleted (loose of information, biasing the results)



- Unconditional mean imputation. Every missing value is substituted by the corresponding global mean of the variable

- Regression imputation. Every missing value is substituted by the predicted value from a multiple regression.

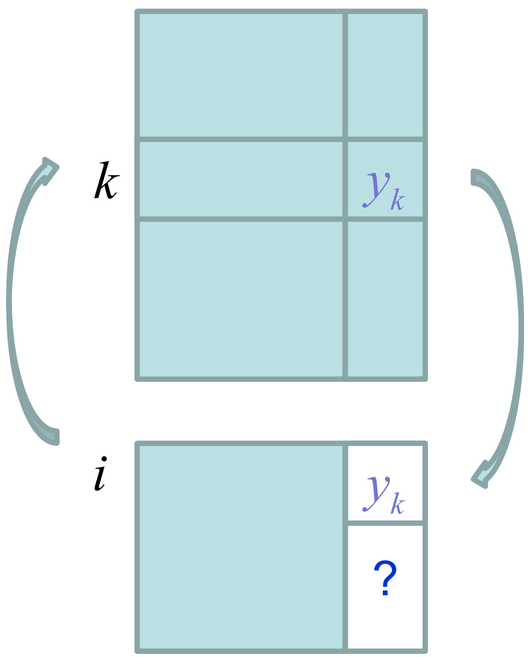


Knn imputation – Method 1

For every observation to be imputed, it identifies ‘k’ closest observations based on the Euclidean distance and computes the weighted average (weighted based on distance) of these ‘k’ obs.

Find the closest individual to i, according all variables except y

Complete data



cases with y missing

Copy the y_k value in the i individual

https://www.r-bloggers.com/2016/04/missing-value-treatment/

Knn imputation (2)

Knn – 1- nearest neighbor imputation (easy to implement)

- For every individual containing a missing value in a specific variable, we find another individual with minimal distance to the previous one with complete information.
- Then transfer (copy) the value of the specific variable, of the second individual to the first one.

https://www.rdocumentation.org/packages/DMwR/versions/0.4.1/topics/knnImputation

Imputation by chained equations -MICE (Method 2)

Let X be a data set with missing observations

Order the data set trying to follow a monotone increasing missingness pattern

1. Start filling in the missing data with values at random
2. For every variable with missing values
   a. Impute the missing values of the variable from the predicted values of the regression of the current variable with the remaining ones.

   Iterate the above procedure till the convergence

   Apply 1nn to impute every missing value from the closest individual to obtain the final realistic imputed values

```
library mice;   imp <- mice(data, m = 1);   data_imp <- complete(imp)
```

https://www.r-bloggers.com/2016/06/handling-missing-data-with-mice-package-a-simple-approach/

Multivariate Imputation By Chained Equations - Example

Step 1: Global Mean or Mode

age	experience	salary(K)	Personal loan
25		50	1
27	3		1
29	5	80	0
31	7	90	0
33	9	100	1
	11	130	0

age	experience	salary(K)
25		50
27	3	
29	5	80
31	7	90
33	9	100
	11	130

age	experience	salary(K)
25	7	50
27	3	90
29	5	80
31	7	90
33	9	100
29	11	130

Step 2: Remove the "age" imputed values and keep the imputed values in other columns as shown here.

age	experience	salary(K)
25	7	50
27	3	90
29	5	80
31	7	90
33	9	100
	11	130

age	experience	salary(K)
25	7	50
27	3	90
29	5	80
31	7	90
33	9	100
	11	130

Step 3: Age was imputed by using Experience and Salary.

age	experience	salary(K)
25		50
27	3	90
29	5	80
31	7	90
33	9	100
34.99	11	130

age	experience	salary(K)
25	0.98	50
27	3	
29	5	80
31	7	90
33	9	100
34.99	11	130

Step 4: "experience" was imputed and proceed now with the last feature, "Salary"

age	experience	salary(K)
25	0.98	50
27	3	70
29	5	80
31	7	90
33	9	100
34.99	11	130

age	experience	salary(K)
25	7	50
27	3	90
29	5	80
31	7	90
33	9	100
29	11	130

minus

age	experience	salary(K)
25	0.98	50
27	3	70
29	5	80
31	7	90
33	9	100
34.99	11	130

age	experience	salary(K)
0	6.02	0
0	0	20
0	0	0
0	0	0
0	0	0
-5.99	0	0

Step 5: Repeat Steps 1 to 4 until reach convergence

Iteration 2

age	experience	salary(K)
25	0.98	50
27	3	70
29	5	80
31	7	90
33	9	100
34.99	11	130

First Dataset

After all imputations

age	experience	salary(K)
25	0.975	50
27	3	70
29	5	80
31	7	90
33	9	100
34.95	11	130

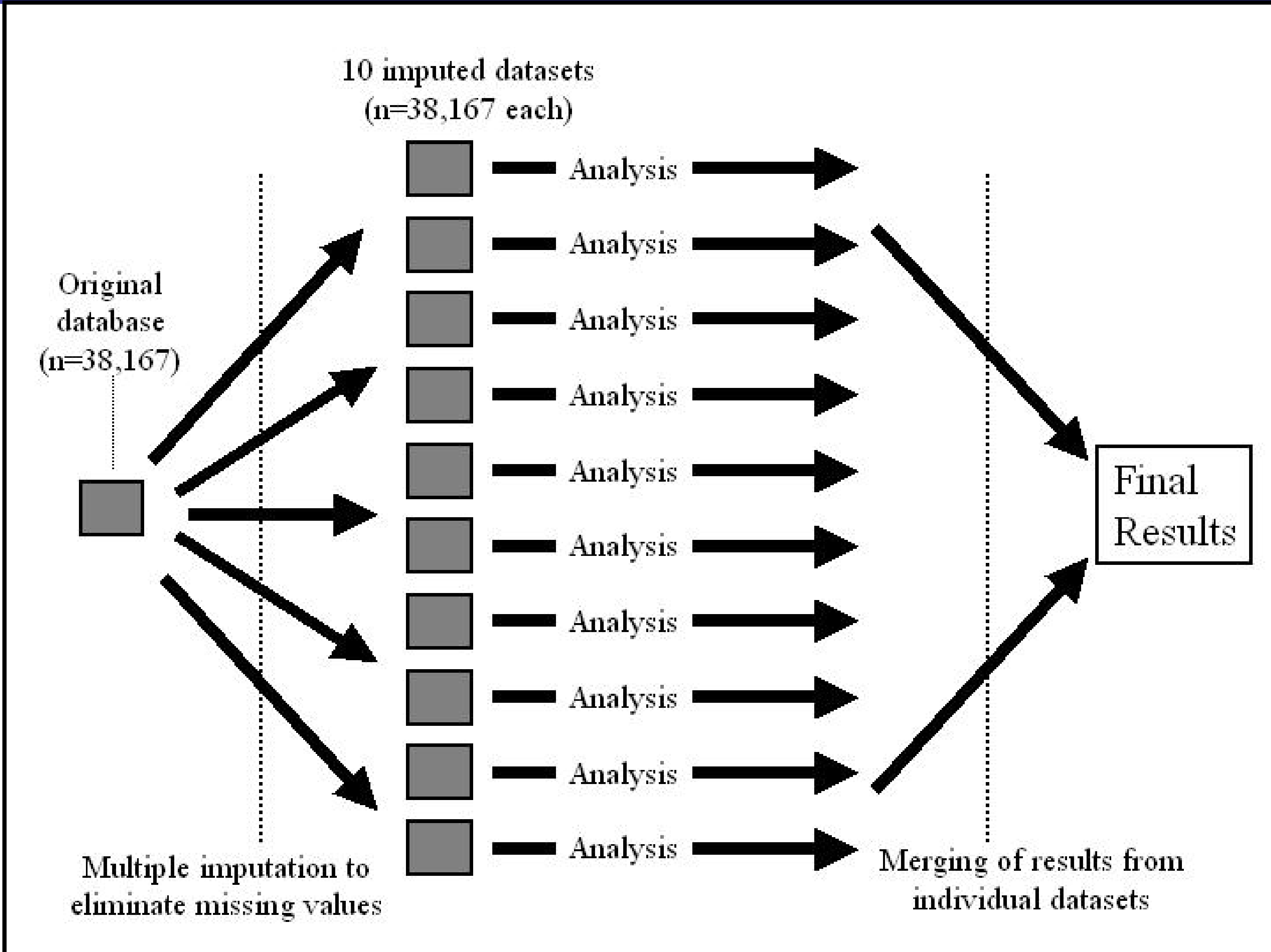
Second Dataset

After Second - First

age	experience	salary(K)
0	0.005	0
0	0	0
0	0	0
0	0	0
0	0	0
0.004	0	0

Difference Matrix

MICE (M Value)


10 imputed datasets
(n=38,167 each)
Original
database
(n=38,167)
Analysis
Analysis
Analysis
Analysis
Analysis
Analysis
Analysis
Analysis
Analysis
Analysis
Final
Results
Multiple imputation to
eliminate missing values
Merging of results from
individual datasets

MICE in R

MICE → Details: try to use in combination with VIM package

```
data<-airquality
tempData <- mice(data,m=5,maxit=50,meth='pmm',seed=500)
summary(tempData)
completedData <- complete(tempData,1)
```

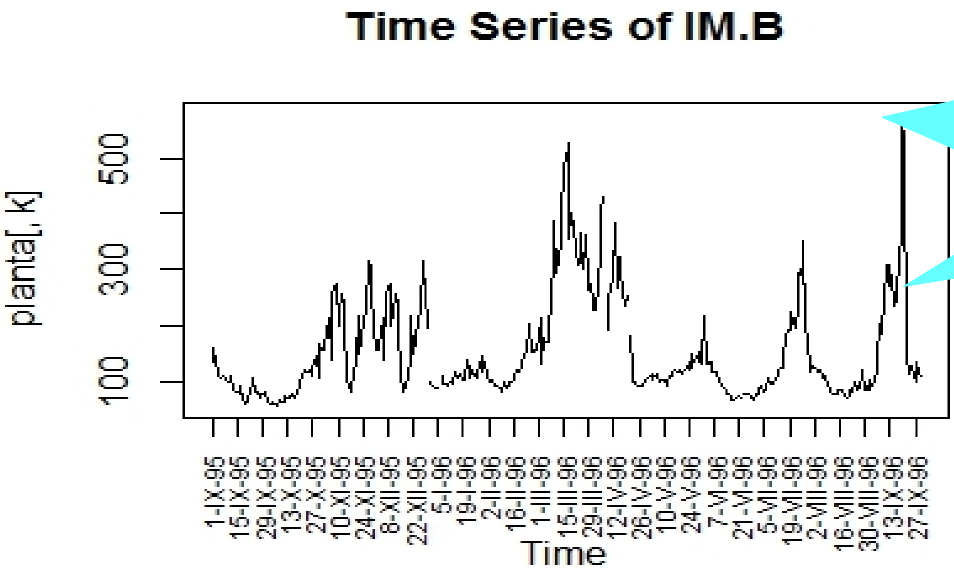
- `m=5` refers to the number of imputed datasets. Five is the default value.
- `meth='pmm'` refers to the imputation method. In this case we are using predictive mean matching as imputation method. Other imputation methods can be used, type `methods(mice)` for a list of the available imputation methods.

Missing Data in Time Series (Method 3)

Interpolation → Usefull for time-series of numerical variables

Linear assumption between observed points

(assume monotonic behaviour between observations)



ALTERNATIVE

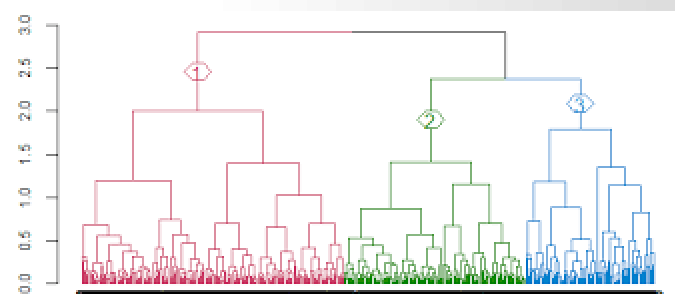
Assume constant between Measurements (slow dynamics)

or Splines for non-linear imputation

Mixed Intelligent-Multivariate Missing Imputation (MIMMI) – Method 4

The MIMMI method [Gibert 2013]

1. Select a small number of relevant variables
2.
3. Use intelligent imputation on that reduced data matrix
4. (expert-based imputation, vertical or horizontal)
5.
6. Multivariate clustering using the imputed variables
7.
8. Impute the missing data of the remaining variables
9. (use mean local to the group of every individual (conditional means)



Imputation by random Forests (Method 5)

Non-parametric method of imputation

Let X be a data set with missing observations of any type (continuous and categorical)

1. Start filling in the missing data with values at random
2. For every variable with missing values

   Impute the missing values of the variable from the predicted values from the random forest of the individuals with the current variable as response using the remaining ones as predictors.

   Iterate the above procedure till imputed values converge

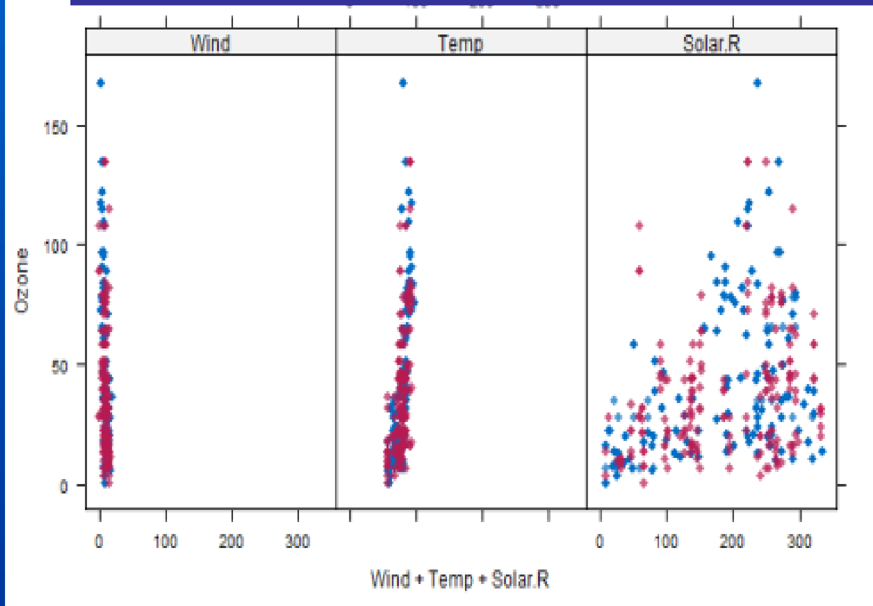
   In convergence, output the OOB error

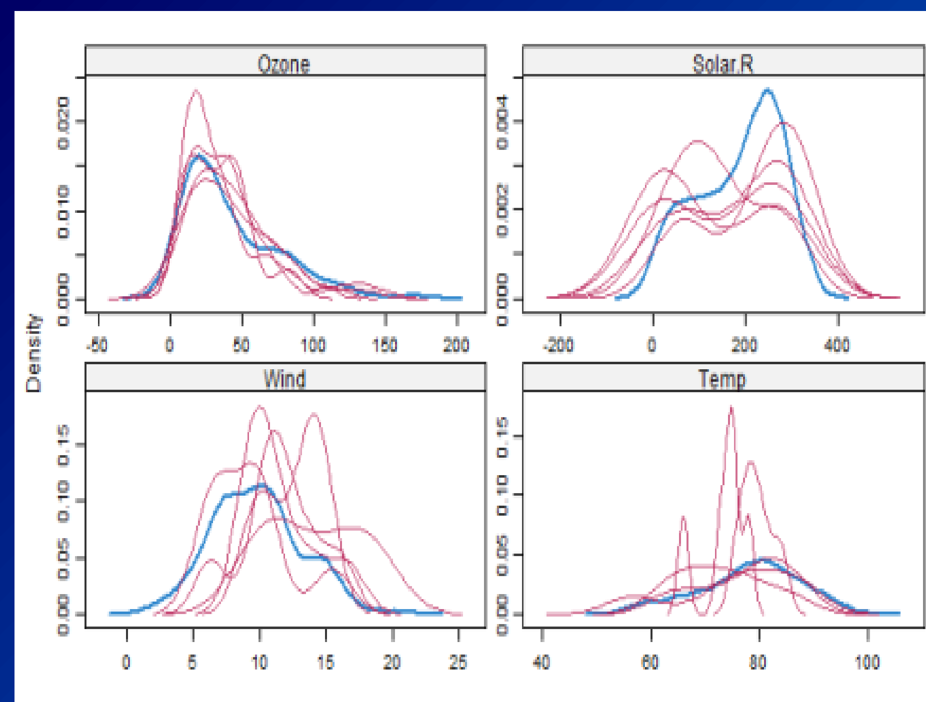
$$\frac{\sum_i (x_{new}^{imp} - x_{old}^{imp})^2}{\sum_i (x_{new}^{imp})^2}$$

```
library(missForest);  mf_imp <- missForest(data); data_imp <- mf_imp$ximp
```

https://rpubs.com/david-deming-tung/missing_dat

Activity # 2 – Comparisson before vs after





Goal # 1 – Summary of methods (Missing Data – Lab Session)

KNN → Suggestion: use previous scripts for Knn Imputation or Library (VIM)
https://cran.r-project.org/web/packages/VIM/vignettes/VIM.html
FROM VIM – 1) Descriptive and Profiling Tools for Missing Data
2) Imputation by Knn
3) Comparisson of results
4) Other useful links:
4.1) https://www.statmethods.net/input/missingdata.html
4.2) http://lib.stat.cmu.edu/R/CRAN/web/packages/mitools/index.html
4.3) https://www.rdocumentation.org/packages/mice/versions/2.25/topics/mice
4.4) Imputation in Time Series
https://cran.r-project.org/web/packages/imputeTS/vignettes/imputeTS-Time-Series-Missing-Value-Imputation-in-R.pdf
5) https://cran.r-project.org/web/views/MissingData.html
(COMPLETE SURVEY for Missing Data Tools in R)

Feature Transformation

1) Data cleaning reasons

Ex. 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, ...

2) Coertions: Information loss.

Discretization (h/week working)

Categorization (Thiroids levels)

Recategorizations (professions)

3)Technical questions:

Estandarditzation, normalitzation and similars

Eventual logaritmic transformation

Required by data mining technique to apply

Feature Selection

- Wrappers:

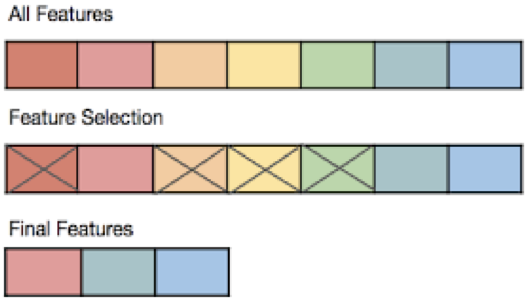
Rank subsets of features by accuracy in predicting Y (costly. Method specific oriented)

- 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)



Feature Extraction

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")

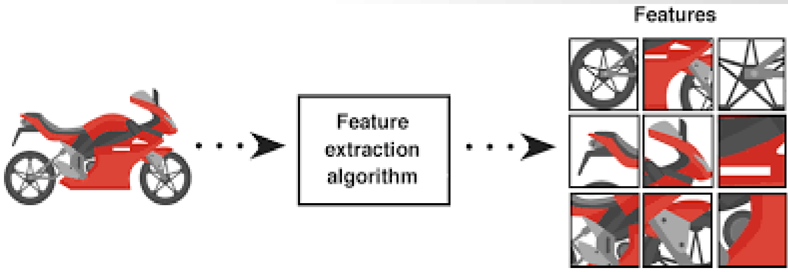
Case Credit Scoring (saving capacity)

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())

S



Goal # 2 – Feature Engineering

FEATURE Transformations → Depends on the model/method

FEATURE SELECTION →

1. Check for previous methods (Statistical Tests, correlations, covariance matrix, PCA)
2. Filter methods pick up the intrinsic properties of the features measured via univariate statistics instead of cross-validation performance. These methods are faster and less computationally expensive than wrapper methods. When dealing with high-dimensional data, it is computationally cheaper to use filter methods (information gain, Chi-square test, Fisher Score & Correlation, etc)
3. Wrappers require some method to search the space of all possible subsets of features, assessing their quality by learning and evaluating a classifier with that feature subset. The feature selection process is based on a specific machine learning algorithm we are trying to fit on a given dataset. It follows a greedy search approach by evaluating all the possible combinations of features against the evaluation criterion. The wrapper methods usually result in better predictive accuracy than filter methods.

FEATURE EXTRACTION → Follow business understanding and domain of the problem