UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Dealing with missing values in modern data science: the good, the bad, and the ugly.

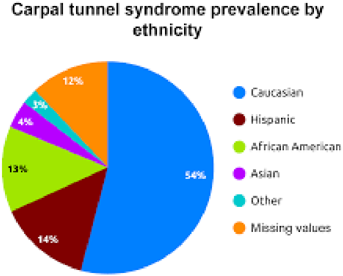
Lluís A. Belanche
belanche@cs.upc.edu

Eleventh European Big Data Management & Analytics Summer School (eBISS 2023)

July 6, 2023

OUTLINE

- Problem
- Incidence
- Causes
- Types
- Methods
- Conclusions
- Opportunities



¿¿

The missing value problem

Missing information is an old issue in statistical analysis! Causes?

- Technical limitations (e.g. sensors working at different measurement rates, only for given periods of time, sensor malfunctioning);
- Measures costly to perform in time or money (e.g., analytical tests) or involving destructive methods (e.g., data from car crash tests);
- Values lost during transmission or storage;
- Senseless values (e.g., number of pregnancies in male adults);
- Reluctance to supply the value (e.g., salaries, phone numbers, etc);
- Many others ...

The missing value problem

Have you ever tried marijuana?

1. Yes: there follow questions specific on marijuana ...
2. No: this specific section is skipped
3. Many researchers would not consider the skipped items to be missing!
4. What about you?
5. It might be a (big) mistake to presume that all skipped items are 0

The missing value problem

Have you ever tried marijuana?

1. Yes: there follow questions specific on marijuana ...
2. No: this specific section is skipped
3. Many researchers would not consider the skipped items to be missing!
4. What about you?
5. It might be a (big) mistake to presume that all skipped items are 0

The missing value problem

What could be the goal?

1. Reconstruct the dataset “as if” it was all complete:
   1. As if it was a plausible dataset from the population
   2. As if it was the original dataset (w/o the lost information)
2. Obtain the same results that we would have seen with complete data (either of the two previous •)
3. Obtain an estimate of the population distribution

What could be the quality metric?

1. Reconstruction error (whatever this means)
2. That of the modelling task
   1. NMSE for regression
   2. GCE for classification
   3. Silhouette index for clustering
   4. ...

The missing value problem

Missing information is difficult to handle! Naïve methods?

1. Remove rows (that is: cases, observations, examples, ...)
2. Remove columns (that is: features, variables, attributes, ...)
3. "Fill in" the hole (*impute*) with the mean, median, mode, ...
4. Nearest neighbour imputation
5. Add another variable equal to one only if the value is absent and zero otherwise
6. Statistical approaches:
    - Parametric ways
    - Non-parametric ways
7. Others??? Sure

The missing value problem

Specially difficult scenarios!

1. When the lost parts are of significant size (say, $> 10\%$);
2. When the missingness pattern is very distributed:
   - Most rows have missing components, AND/OR
   - Most columns have missing components
3. When the missingness pattern has many (unknown) dependencies;
4. Others? Try yourself ...

The missing value problem

There are three big ways of dealing with missing data:

1. **Discard** the involved data until the data is complete;
2. **Impute** the involved data in a hopefully "optimal" way;
3. **Extend** the methods to be able to work with incomplete data

The missing value problem

- When missing values occur for reasons beyond our control, we must make assumptions about the processes that create them;
- These assumptions are usually untestable;
- Good science suggests that assumptions be made explicit!

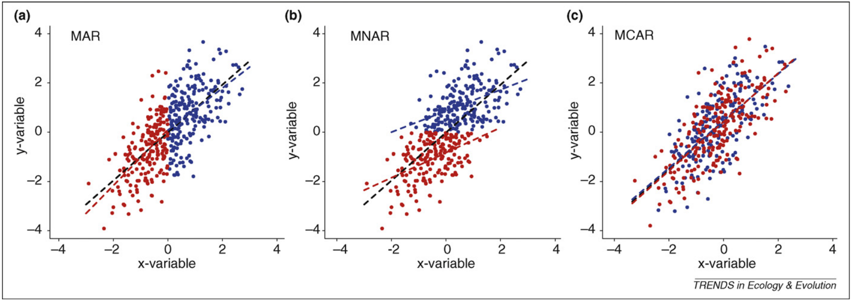
The three main distributions of missing data

1. Missingness completely at random (MCAR)
2. Missingness at random (MAR)
3. Missingness not at random (MNAR)

These terms are defined by theoretical relationships between missing values and observed or unobserved variables.

The missing value problem

(from Nakagawa S, Freckleton RP. Missing in action: the dangers of ignoring missing data. Trends Ecol Evol. 2008 Nov;23(11):592-6.)



What value should I use to indicate missing data?

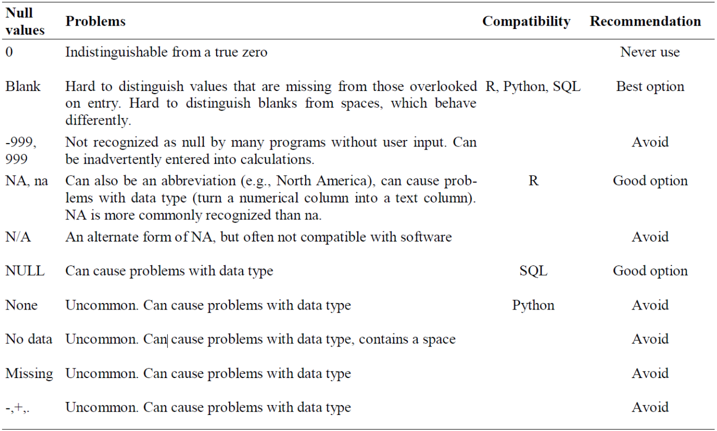
Use a value that is:

1. Compatible with (most) software;
2. Coherent across all data types;
3. Unlikely to cause errors in storage or analysis.

Use a value that is:

1. Convenience: liberates the programmer from a cumbersome task;
2. Orthogonality: it works across most data types and methods;
3. Minimises the chances of introducing semantic errors in the program

What value should I use to indicate missing data?



Null values	Problems	Compatibility	Recommendation
0	Indistinguishable from a true zero		Never use
Blank	Hard to distinguish values that are missing from those overlooked on entry. Hard to distinguish blanks from spaces, which behave differently.	R, Python, SQL	Best option
-999, 999	Not recognized as null by many programs without user input. Can be inadvertently entered into calculations.		Avoid
NA, na	Can also be an abbreviation (e.g., North America), can cause problems with data type (turn a numerical column into a text column). NA is more commonly recognized than na.	R	Good option
N/A	An alternate form of NA, but often not compatible with software		Avoid
NULL	Can cause problems with data type	SQL	Good option
None	Uncommon. Can cause problems with data type	Python	Avoid
No data	Uncommon. Can cause problems with data type, contains a space		Avoid
Missing	Uncommon. Can cause problems with data type		Avoid
-,+,.	Uncommon. Can cause problems with data type		Avoid

(from Ideas in Ecology and Evolution 6(2): 1–10, 2013)

1: Discarding data

Discarding data

- The possibility of simply discard the involved data can not be considered as a "method" and is also frustrating because of the lost effort in collecting the information;
- It also causes a decrease in statistical significance
- This can be done only if the number of missing values is very small or else they are heavily concentrated in some variables.
- However, it creates an interesting research problem: (rows and/or columns)

2: Imputing data

The good, the bad, and the ugly!

We can classify representative imputation methods according to a number of quality metrics:

- simplicity/elegance
- correctness/soundness
- flexibility/assumptions
- interpretability
- computational effort

2: Imputing data

A selection of classical and modern methods:

- Impute with the mean, median, mode, ...
- Dummy variable adjustment
- Nearest neighbour imputation
- Maximum likelihood imputation
- Expectation maximization imputation
- Markov chain Monte Carlo imputation
- MissForest
- MICE
- Bayesian multiple imputation

The missing value problem

Which methods (willingly) accept missing values?

- Nearest neighbours How?
- Decision trees How?
- Naïve Bayes How?
- Support Vector Machines? How?
- others? ... an interesting research problem (adapt)

The missing value problem

Which methods (willingly) accept missing values?

- Nearest neighbours How?
- Decision trees How?
- Naïve Bayes How?
- Support Vector Machines? How?
- others? ... an interesting research problem (adapt)

The missing value problem

Which methods (willingly) accept missing values?

- Nearest neighbours How?
- Decision trees How?
- Naïve Bayes How?
- Support Vector Machines? How?
- others? ... an interesting research problem (adapt)

The missing value problem

Which methods (willingly) accept missing values?

- Nearest neighbours How?
- Decision trees How?
- Naïve Bayes How?
- Support Vector Machines? How?
- others? ... an interesting research problem (adapt)

The missing value problem

Which methods (willingly) accept missing values?

- Nearest neighbours How?
- Decision trees How?
- Naïve Bayes How?
- Support Vector Machines? How?
- others? ... an interesting research problem (adapt)

The missing value problem

Which methods (willingly) accept missing values?

- Nearest neighbours How?
- Decision trees How?
- Naïve Bayes How?
- Support Vector Machines? How?
- others? ... an interesting research problem (adapt)

The missing value problem

Which methods (willingly) accept missing values?

- Nearest neighbours How?
- Decision trees How?
- Naïve Bayes How?
- Support Vector Machines? How?
- others? ... an interesting research problem (adapt)

The missing value problem

Which methods (willingly) accept missing values?

- Nearest neighbours How?
- Decision trees How?
- Naïve Bayes How?
- Support Vector Machines? How?
- others? ... an interesting research problem (adapt)

The missing value problem

Which methods (willingly) accept missing values?

- Nearest neighbours How?
- Decision trees How?
- Naïve Bayes How?
- Support Vector Machines? How?
- others? ... an interesting research problem (adapt)

Example in a real application domain

Handling missing values in microbiology

- The study of fecal source pollution in waterbodies is a major problem in ensuring the welfare of human populations
- **Microbial source tracking** (MST) methods attempt to identify the source of contamination, allowing for improved risk analysis and better water management
- The available dataset includes 148 observations about 10 chemical, microbial, and eukaryotic markers of fecal pollution in water
- All variables (except the class variable) are binary, i.e., they signal the presence or absence of a particular marker

Example in a real application domain

Origin	HF183	HF134	CF128	Humito	Pomito	Bomito	ADO	DE
Human :50	0 :68	0 :81	0 :104	0 :35	0 :83	0 :78	0 :56	0 :
Cow :26	1 :40	1 :26	1 : 5	1 :79	1 :32	1 :32	1 :59	1 :
Poultry:31	**?**:31	**?**:32	**?**:30	**?**:25	**?**:24	**?**:29	**?**:24	**?**:2
Pig :32								

- Summary (counts) table for the full dataset. The first column is the target class.
- The symbol **?** denotes a missing value.
- The percentage of missing values is around 19.8%, and all the predictive variables have percentages between 17% and 23%

Example in a real application domain

Theorem. Let the symbol $\mathbf{?}$ denote a missing element, for which only equality is defined, and $\mathcal{X}$ a finite discrete set. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel in $\mathcal{X}$ and P a probability mass function in $\mathcal{X}$. Then the function $k^{\mathbf{?}}(x, y)$ given by

$$k^{\mathbf{?}}(x,y) := \begin{cases} k(x,y), & \text{if } x, y \neq \mathbf{?}\ ; \\ g(x) := \sum\limits_{y' \in X} P(y')k(x,y'), & \text{if } x \neq \mathbf{?} \text{ and } y = \mathbf{?}; \\ g(y) := \sum\limits_{x' \in X} P(x')k(x',y), & \text{if } x = \mathbf{?} \text{ and } y \neq \mathbf{?}; \\ G := \sum\limits_{x' \in X} P(x') \sum\limits_{y' \in X} P(y')k(x',y'), & \text{if } x = y = \mathbf{?} \end{cases}$$

is a kernel in $\mathcal{X} \cup \{\mathbf{?}\}$.

Example in a real application domain

For the particular case of binary variables $x, y \in \{v_1, v_2\}$, a convenient approach is to define the kernel:

$$k_{0/1}(x, y) := \mathbb{I}_{\{x=y\}}$$

where

$$\mathbb{I}_{\{z\}} = \begin{cases} 1 & \text{if } z \text{ is true} \\ 0 & \text{if } z \text{ is false} \end{cases}$$

Example in a real application domain

Consider now $\boldsymbol{x}, \boldsymbol{y} \in \{0, 1, \textbf{?}\}^p$. When we apply the Theorem to this kernel, we obtain an extended multivariate kernel:

$$\mathcal{K}_1(\boldsymbol{x}, \boldsymbol{y}) := \frac{1}{p} \sum_{i=1}^{p} \begin{cases} 1 & \text{if } x_i = y_i = 1 \text{ ;} \\ P_i(x_i), & \text{if } x_i \neq \textbf{?} \text{ and } y_i = \textbf{?}; \\ P_i(y_i), & \text{if } x_i = \textbf{?} \text{ and } y_i \neq \textbf{?}; \\ (P_i(0))^2 + (P_i(1))^2, & \text{if } x_i = y_i = \textbf{?}; \\ 0, & \text{otherwise} \end{cases}$$

This kernel is a generalization of the classical *Simple Matching Coefficient*, proposed by Sokal and Michener for numerical taxonomy

Example in a real application domain

Alternatives???
Given $\boldsymbol{x}, \boldsymbol{y} \in$, let $c(\boldsymbol{x})$ be the set of completions of $\boldsymbol{x}$. Given two vectors $\boldsymbol{x}, \boldsymbol{y} \in \{0, 1, \mathbf{?}\}^p$, the function

$$\mathcal{K}_2(\boldsymbol{x}, \boldsymbol{y}) := \frac{1}{|c(\boldsymbol{x})||c(\boldsymbol{y})|} \sum_{\boldsymbol{x}' \in c(\boldsymbol{x})} \sum_{\boldsymbol{y}' \in c(\boldsymbol{y})} k(\boldsymbol{x}', \boldsymbol{y}')$$

is a kernel in $\{0, 1, \mathbf{?}\}^p$.

Example in a real application domain

			10x10cv for each class			
Approach	C	10x10cv	Human	Cow	Poultry	Swine
$\mathcal{K}_1$	2.0	79.3	95.4	64.5	75.2	69.4
$\mathcal{K}_2$	1.6	78.2	92.6	62.8	71.8	74.2
MI-1	1.0	79.9	92.7	66.4	69.4	80.2
MI-2	1.0	79.0	94.5	57.5	70.8	78.8

Mean 10x10cv accuracies for the four approaches to handle missing values. Also shown are best cost parameter C and detailed class performance.

(*joint work with G. Nebot, T. Aluja and V. Kobayashi*)

Example: the Pima Indian Diabetes Dataset

- The Pima Indian Diabetes Dataset, originally from the National Institute of Diabetes and Digestive and Kidney Diseases, contains information of 768 women from a population near Phoenix, Arizona, USA.
- The purpose is to study the associations between having diabetes and various physiological characteristics. Although there are surely other factors (including genetic) that influence the chance of having diabetes, the hope is that by having women who are genetically similar (all from the Pima tribe), that these other factors are naturally accounted for.

Example: the Pima Indian Diabetes Dataset

```
Number of times pregnant
Plasma glucose concentration (glucose tolerance test)
Diastolic blood pressure (mm Hg)
Triceps skin fold thickness (mm)
2-Hour serum insulin (mu U/ml)
Body mass index (weight in kg/(height in m)^2)
Diabetes pedigree function
Age (years)
```

Watch out!!!

- While the UCI repository index claims that there are no missing values, closer inspection of the data shows several physical impossibilities, e.g., blood pressure or body mass index of 0!
- All zero values of glucose, pressure, triceps, insulin and mass should be set to NA.

The missing value problem

Example of *ad hoc* (tricky) method for CONT data

Suppose we replace the missing values by an arbitrary number (say, 0) and introduce a dummy indicator X_{miss} that is 1 if X is missing and 0 if X is observed.

This procedure merely redefines the coefficients

- In the original model $\mathbb{E}(Y) = \beta_0 + \beta_1 x$, where β_0, β_1 represent the intercept and slope for the full population;
- In the expanded model $\mathbb{E}(Y) = \beta_0 + \beta_1 x + \beta_2 x$, where β_0, β_1 represent the intercept and slope for respondents, and β_0, β_2 represent the mean of Y among nonrespondents.

MAR, MCAR and MNAR

Example of *ad hoc* (tricky) method for CATEG data

Suppose that missing values occur on a nominal outcome with response modalities $1, 2, \ldots, k$.
Suppose we replace the missing values by introducing a new modality (say, $k + 1$).

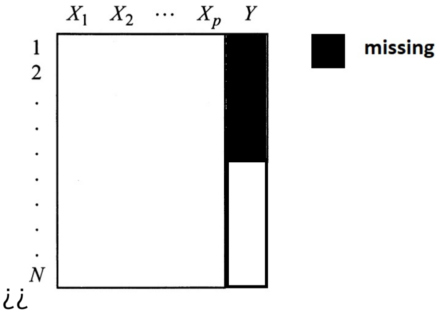
This procedure merely redefines the coefficients

Again, it merely redefines modalities $1, 2, \ldots, k$ so to apply to respondents only.

The missing value problem

Today, missingness is regarded as a probabilistic phenomenon: it has a distribution: assume we have a set of i.i.d. measured variables $(X_1, \ldots, X_p, Y)$ and Y can have missing values.

- [MCAR] The prob. that Y_0 is missing does not depend on the values of Y_0, X_0;
- [MAR] The prob. that Y_0 is missing depend on the values of X_0 but not on Y_0;
- [MNAR] The prob. that Y_0 is missing may depend on the value of Y_0 itself.



Utilities

The field is huge and it is not possible to cover all methods or ideas here; the interested practitioner is referred to:

```
[books] MISSING DATA: A GENTLE INTRODUCTION.
        Patrick E. McKnight,
        Katherine M. McKnight,
        Souraya Sidani,
        Aurelio Figueredo
[tutorials]
    Schafer, J. L., Graham, J. W. (2002). Missing data:
    Our view of the state of the art.
    Psychological Methods, 7(2), 147{177
[web resources]
    www.missingdata.nl
    rmisstastic.netlify.app/
    www.asc.ohio-state.edu/kaizar.1/courses/652/
    gking.harvard.edu/amelia
[software]
    cran.r-project.org/web/views/MissingData.html
```

Conclusions

- Why do missing data create such difficulty in scientific research? Because most data analysis methods are not designed for them;
- *Ad hoc* edits may do more harm than; good, producing answers that are biased, inefficient (lacking in power), and unreliable;
- When the rate of missing values is high, the chosen method will exert a high degree of influence over the results, and differences among competing methods will be magnified;
- When missingness is beyond the researcher's control, its distribution is unknown and MAR is only an assumption. In general, there is no way to test whether MAR holds in a data set.

Laboratory work

Do-it-yourself!

- Pima Indians dataset
- Explore LP possibilities, implications, warnings, ...
- Create a synthetic dataset and play (methods, %, sample size, ...)
- Choose a dataset of your interest
- Think of other interesting research problems