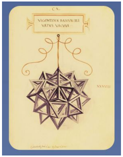
The use of multiple imputation techniques in social media data

Paolo Mariani, Andrea Marletta, Matteo Locci

University of Milano-Bicocca



Cladag 2021
13-th Scientific Meeting Classification and Data Analysis Group
September 9-11, 2021

1. Brief introduction to missing values in social data
2. Imputation methods: Threshold value rule and MIMCA approach
3. Application and results
4. Summary and Conclusions

Research objectives

The aims of this study:

- to offer an innovative analysis path to discern the missing value from a behaviour
- to consider proposed imputation methods as a data enrichment technique
- to apply this technical issue to social media data seen as source of incompleteness

Social data represent a source of uncertainty because of their collection method, besides robust statistical techniques are necessary to reduce missing data uncertainty.

Missing value framework

Missing values = lack of information

In not statistical contexts, row elimination is the most abused solution. This method consists in the deletion of each observation containing missing values, but it may be misleading because using this method the lack of information is increased.

Statistically speaking, it is better to understand and recognise the nature and the mechanism underlying the lack of information evaluating the link between the observed value and the missing one.

Well-known classification of Little and Rubin:

- MCAR (Missing Completely At Random)
- MAR (Missing at Random)
- MNAR (Missing Not At Random)

Imputation techniques

If data are not MCAR, then the presence of behaviour associated to the missing values could be hypothesized. A two-steps procedure was implemented to impute the missing values:

1. First substitution of missing values was obtained using a threshold based on the number of present values in the sparse matrix.
2. Second substitution through a multiple imputation technique known as the MIMCA (Multiple Imputation with Multiple Correspondence Analysis) method (Audigier et al., 2017).

This approach could be seen as a path of data enrichment using the initial database to enrich data, assuming that the missing observations could be the result of a behaviour

Threshold value rule

Let consider a sparse matrix with 0 and 1 where 0 values represent a missing observation and 1 values an appreciation for a variable, this method is based on the hypothesis that, the higher the occurrence of 1 values, the lower the probability that 0 values may be imputable as missing observation.

Let define l_i, for $i = 1, \ldots, n$, and $l_i \in (0, 1]$ as the proportion of the total number of 1 values for each considered variable divided for the total number of variables.

The choice of the threshold for the disambiguation is determined from the average value of the proportion of 1 values for each row. Formally, let c the threshold, it could be defined as:

$$c = \frac{1}{n} \sum_{i=1}^{n} l_i .$$

Threshold value rule (2)

In particular, the decisional rule for the disambiguation is:

- if $I_i > c$, it is supposed that the i-th observation is interested in the content of the variables. Operationally, the 0 values are imputed as an expression of a negative opinion of the content of the variables
- if $I_i < c$, it is supposed that the i-th observation is not interested in the content of the variables. Therefore, the 0 values are not imputed as an expression of a negative opinion of the content of the variables but they are still missing values;
- if $I_i \approx c$, further information is necessary to investigate the behaviour of the i-th observation and additional control is needed to solve the uncertainty. For the moment, the 0 values are still not imputed.

The MIMCA approach

For the statistical units with a proportion I_i very close to the threshold, an available alternative consists in introducing a second imputation technique.

Multiple imputation using MCA allows to impute datasets with incomplete categorical variables. The principle of MI with MCA, consists in creating M different datasets obtained with an algorithm called *iterative MCA*.

This algorithm consists in recoding the incomplete data set, estimating the principal components and loadings from the completed matrix and then, using these estimates to impute missing values.

After the first step of imputation, the procedure of iterative MCA is repeated many times until a convergence criterion is reached.

The MIMCA approach (2)

MIMCA approach is based on regularised iterative MCA. In order to consider the uncertainty concerning the imputed values, M data sets are created according to a bootstrap approach.

The number of components is chosen with a repeated cross-validation. Imputing M datasets, a threshold d is assumed to define the decision rule:

- if the proportion of the imputation as negative opinion is higher than $50\% + d$, then the missing value is imputed as an expression of negative opinion;
- if the proportion of the imputation as negative opinion is lower than $50\% - d$, then the missing value is imputed as an expression of neutral behaviour;
- if the proportion of the imputation as negative opinion is between $50\% - d$ and $50\% + d$, then the missing value is not imputed.

Application

- In the context of social network, a form of interaction is represented by an active presence using some specific tools as a ”Like”
- Placing a ”Like”, users show approval for a content stating a preference and showing positive feedback
- In order to analyze the role of the ”Likes” in this context, a dataset has been considerer containing information for $1,000$ Italian subjects that expressed at least one “Like” for a set of social media pages, websites, and forums concerning healthcare
- The selected category for social network pages is Italian newspapers. Each column of the dataset is a dummy variable that represents the presence or absence of a “Like.”

Application (2)

The 7 Italian newspapers are:

- La Repubblica
- Corriere della Sera
- Il Fatto Quotidiano
- Il Sole 24 Ore
- La Gazzetta dello Sport
- Il Messaggero
- La Stampa

Number of "Like"	0	1	2	3	4	5	6	7
Frequency	504	174	127	78	60	24	16	17

Frequency distributions

"Like" distribution for social page

	La Repubblica	**Corriere della Sera**	**Il Fatto Quotidiano**	**Il Sole 24 Ore**
"Like"	299	244	268	158
Missing Values	197	252	228	338
Missing Values (%)	39.7	50.8	46	68.1
	La Gazzetta dello Sport	**Il Messaggero**	**La Stampa**	**Total**
"Like"	116	66	86	1237
Missing Values	380	430	410	2235
Missing Values (%)	76.6	86.7	82.7	64.7

Joint "Like" distribution for user and social page

"Like"	**La Repubblica**	**Corriere della Sera**	**Il Fatto Quotidiano**	**Il Sole 24 Ore**	**La Gazzetta dello Sport**	**Il Messaggero**	**La Stampa**
1	46	34	51	15	19	5	4
2	84	48	72	20	20	6	4
3	64	52	47	19	25	5	12
4	50	54	45	39	18	14	20
5	23	23	20	22	10	8	14
6	15	16	16	16	7	11	15
7	17	17	17	17	17	17	17

Imputation with threshold rule

If a user placed 1 "Like" over 7 pages, therefore the proportion of observed "Like" is $l = 0.14$. Therefore, for this application $l_i \in [0.14, 1]$, for $i = 1, \ldots, n$. After repeating for all the users, it is possible to extract the average of the proportions of observed "Likes", obtaining the threshold value for the disambiguation $c = 0.36$.

The decisional rule for the disambiguation is specified as follows:

- if $l_i < 0.36$, the missing value will be imputed with a “Nothing” assuming that the user does not know the other social pages
- if $l_i > 0.36$, the missing value will be imputed with a “Dislike” assuming that the user know the social pages
- if $l_i \approx 0.36$, the missing value will be not imputed. The available information is not sufficient to make clear disambiguation.

Imputation with threshold rule (2)

Results of the first imputation method showed that 1288 missing values have been imputed, 244 “Dislike” and 1044 “Nothing”. There are still 947 not disambiguated missing values, but the percentage of missing values has been decreased from 64.7% to 27.8%.

	la Repubblica	**Corriere della Sera**	**Il Fatto Quotidiano**	**Il Sole 24 Ore**
“Like”	299	244	268	158
“Dislike”	12	7	19	23
“Nothing”	128	140	123	159
Missing Values	57	105	86	156
	La Gazzetta dello Sport	**Il Messaggero**	**La Stampa**	**Total**
“Like”	116	66	86	1237
“Dislike”	65	67	51	244
“Nothing”	155	169	170	1044
Missing Values	160	194	189	947

Imputation with MIMCA

The process of data enrichment about cells without a category observed or imputed can be completed using MIMCA approach. Different scenarios with $M = 25, 50, 75, 100$ datasets have been imputed with MIMCA. Best results in prediction terms have been obtained for $M = 100$.

To minimize the simulation error due to the application of a bootstrap procedure, a d threshold has been introduced equal to 10% obtaining the following imputation rule:

- if the proportion of “Dislike” imputed is greater than or equal to 60%, then a “Dislike” is imputed
- if the proportion of “Dislike” imputed is less than or equal to 40%, then a “Nothing” is imputed
- if the proportion of “Dislike” is between 40% and 60%, then neither a “Dislike” nor a “Nothing” is imputed

Imputation with MIMCA (2)

	La Repubblica	Corriere della Sera	Il Fatto Quotidiano	Il Sole 24 Ore	La Gazzetta dello Sport	Il Messaggero	La Stampa	Total
“Like”	299	244	268	158	116	66	86	1237
“Dislike”	24	8	57	47	221	255	200	812
“Nothing”	149	235	139	197	155	169	170	1214
Missing Values	24	9	32	94	4	6	40	209

As can be noted from the table, few missing values are still present with a proportion of missing values now equal to 6.4%.

Conclusions

- This study proposed an alternative to enrich a sparse matrix made of 0 and 1 using a data-driven approach based on multiple imputation techniques
- The two imputation methods created decisional rules founded respectively on a threshold value and MIMCA approach
- In a social media context, in which values 1 stands for a presence of "Like", this approach could help to distinguish the missing value as a neutral or negative behaviour
- In the presented application on Italian social pages about newspapers, the proportion of missing values has been considerably reduced
- Future works could regard similar analysis conducted on different datasets to generalize the method and verify the goodness and the applicability of the model