Pre-processing of animal feed data: an essential step

F. Maroto, A. Gómez, J.E. Guerrero, A. Garrido, D. Sauvant, G. Tran, V. Heuzé and D.C. Pérez

Introduction

- Feed laboratories and research centres generate countless data of chemical composition and nutritive value for specific research purposes or for quality control.
- These data can be useful for data mining purposes, such as building feed tables or creating predictive equations.
- However, real-world data tend to be heterogeneous, noisy, inconsistent and incomplete.
- **Pre-processing**, and particularly the handling of **outliers** and **missing data**, is necessary in order to improve the suitability of feed data for their subsequent analysis

One dataset, two studies

- The database includes about 19,000 samples of alfalfa (*Medicago sativa* L.)
  - Fresh, hay, dehydrated, silage
  - 21 descriptive metadata (process, origin, variety, year, maturity, age, cut…)
  - 25 chemical and nutritive attributes (proximate analysis, minerals, *in vitro* and *in vivo* digestibility…)

- Sources
  - 217 scientific papers
  - 13 databases (Spain, France, North Africa…)

Metadata issues

- There is a considerable lack of uniformity in feed metadata
  - Synonyms
    - Pelleted, granulated
  - Homonyms
    - « First cut » is the cutting carried out for weed control, or the first usable harvest
  - Overlapping and/or ambiguous concepts
    - Terms that describe age and/or maturity still vary widely in the literature
  - General need for a feed-specific domain ontology

Outliers study

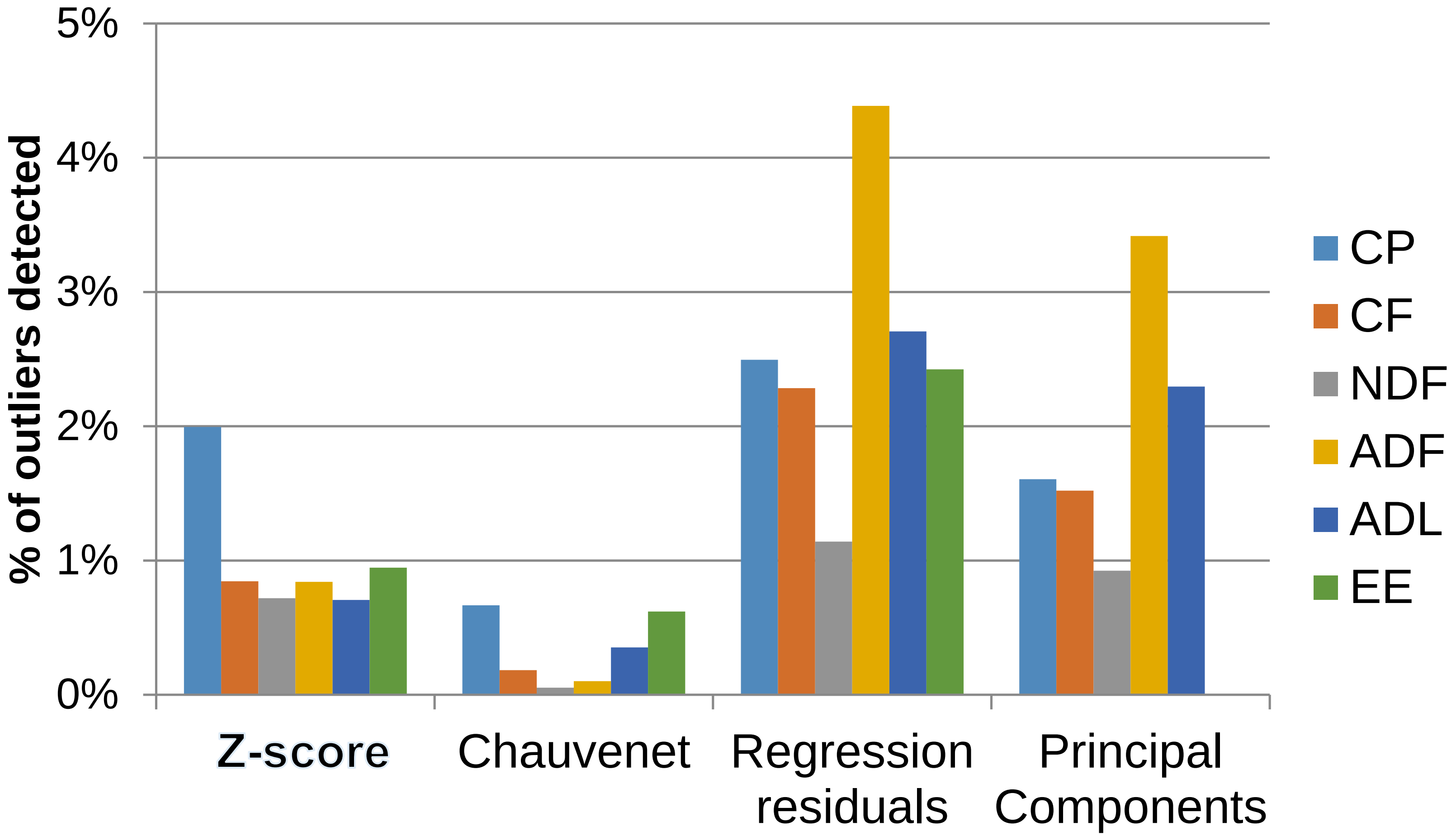
- Several methods for detecting outliers were compared :

Univariate	Bivariate	Multivariate
Z-score Criterion: z = • x – ¼/ Ã• e 3	**Regression residuals** Linear regressions with pairs of variables and Z-score criterion to studentized residuals	**Adjusted Wilks** d_m^2 approximated to a Snedecor f value Criterion: d_m^2 e 3
Chauvenet's criterion P = probability that the data point furthest from the mean has the value assigned by the normal distribution Citerion: P x n d 0,5	**Principal Components** PCs with pairs of variables and Z-score criterion to PC2	**Local Outlier Factor** Compares the local density of a point with the density of its neighbours (N=100) Criterion: LOF > 2 (normality not required)

Outliers detection for univariate and bivariate methods

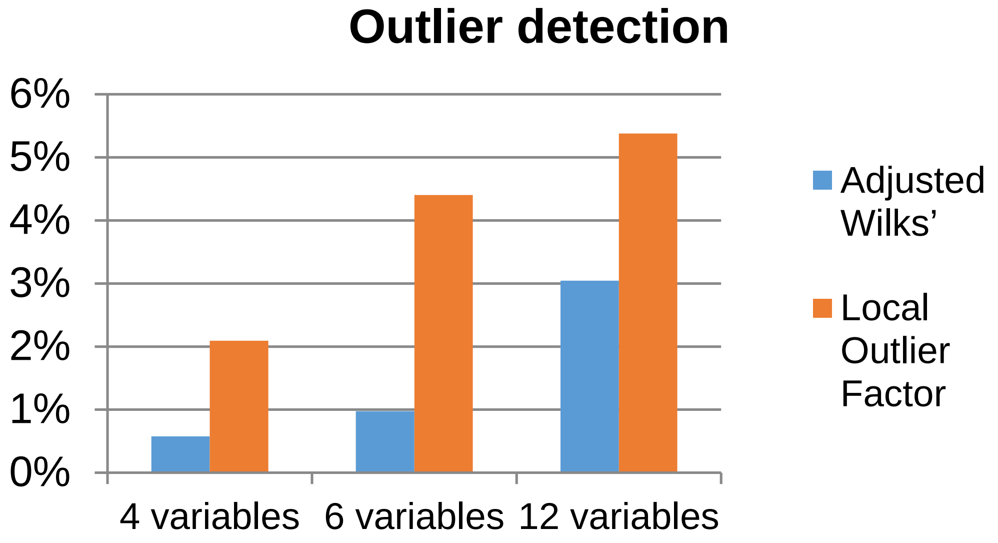
- Univariate methods
  - Z-score > Chauvenet's criterion
  - Many false positives for DM (Z-score)
    - It is necessary to take into account metadata
- Bivariate methods
  - Regression residuals > Principal Components
  - Availability depends on the relations between parameters
    - **CP, CF, NDF, ADF, Lignin, Ca**: 90-100% data can be tested
    - **Ash, Na**: 40-50% of the data
    - **DM, EE**: < 5% of the data due to poor correlations with other parameters

Outlier detection for univariate and bivariate methods

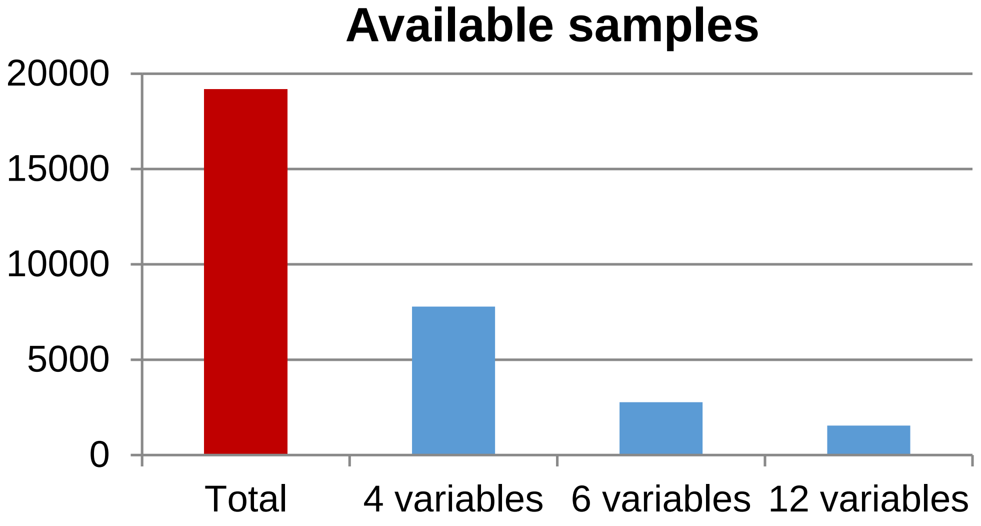


Outlier detection for multivariate methods

- LOF > Adjusted Wilks
- LOF finds outliers not detected by other methods



- Loss of samples increased with the number of variables taken into account



Qualitative characterization of outliers

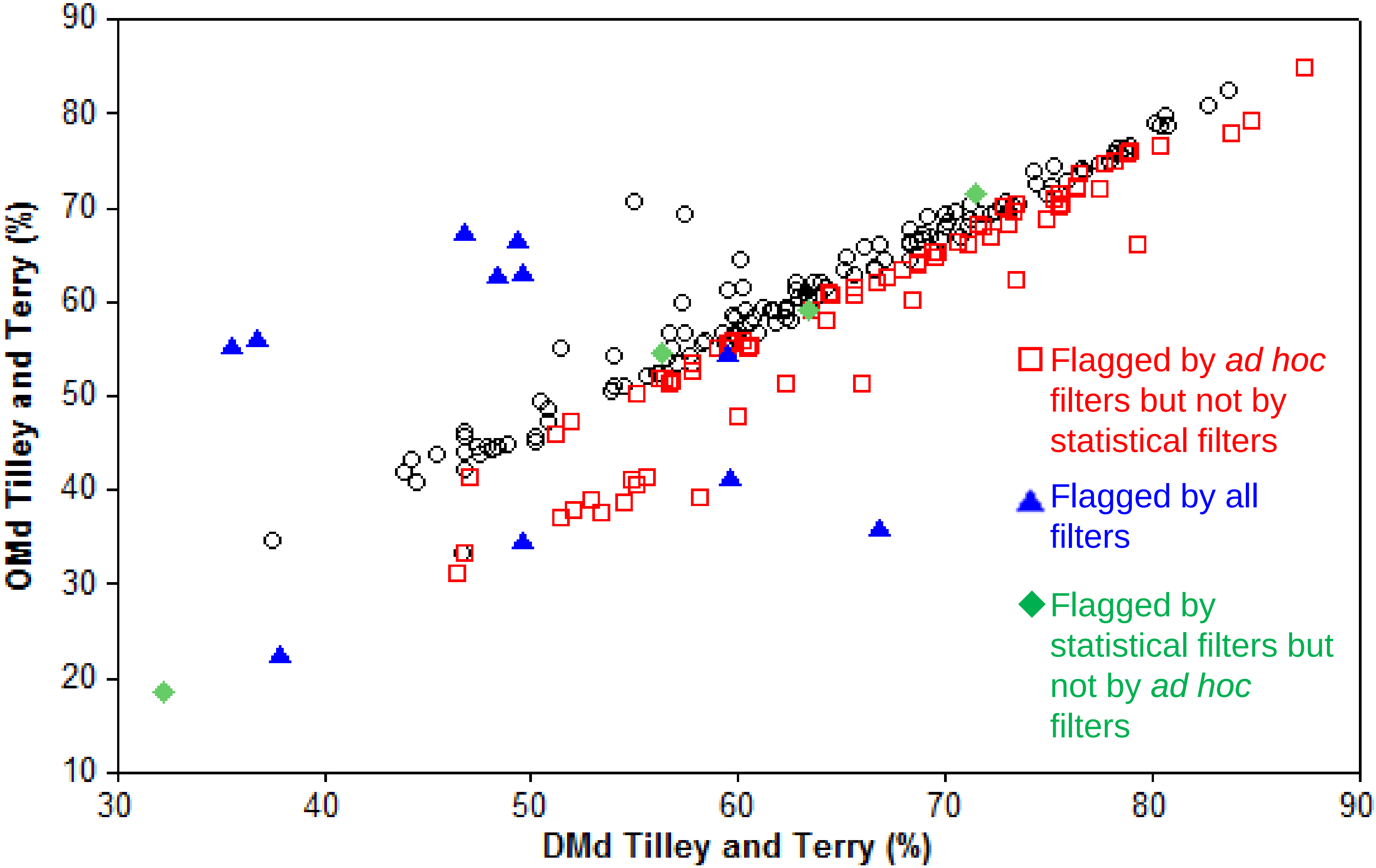
- Transcription errors
  - Example: misplaced decimal point
- Interpretation errors
  - *In vitro* measurements mistaken for *in vivo* ones
- Analytical issues
  - Contamination by soil ➔ high ash values
- Uncommon values
  - Very mature samples, urea-treated silage

Utilization of *ad hoc* filters

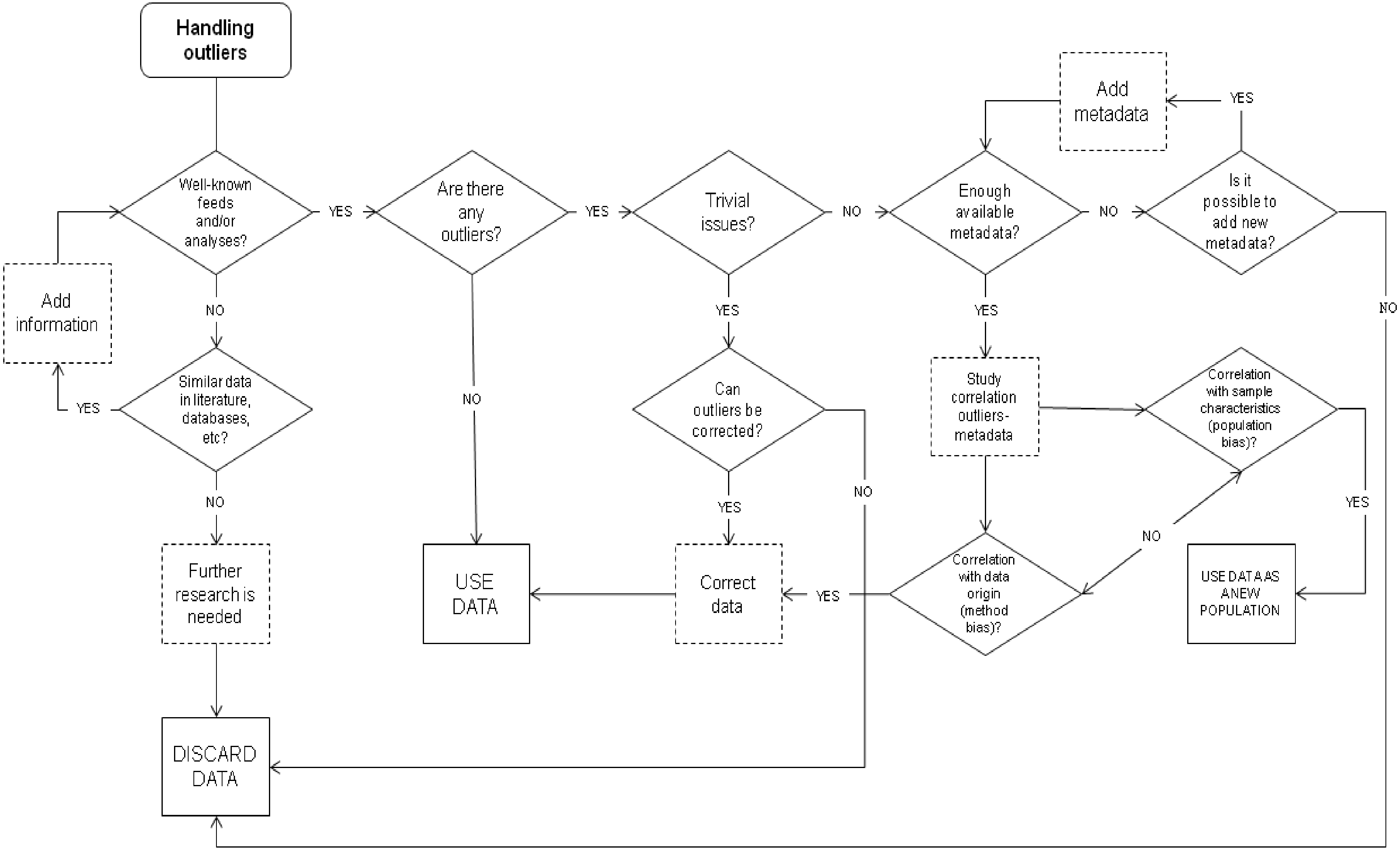
- Statistical filters cannot detect all kinds of outliers: ad hoc filters are necessary

Ad hoc filters	Errors
NDF > ADF	9
ADF > ADL	0
NDF > ADL	0
ADF > CF	59
ASH > • Minerals	1
OMD *in vivo* > DMD *in vivo*	9
OMD *in vitro* – DMD *in vitro* between DMD x (Ash/OM) and DMD x (Ash/OM) –100 x (Ash/OM)	147

Statistical filters *vs. ad hoc* filters



Heuristic approach

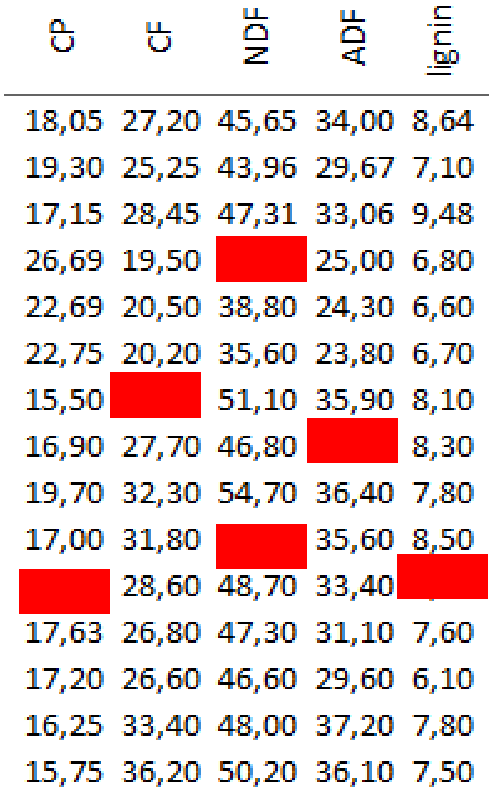

Handling outliers
Well-known feeds and/or analyses?
YES
Are there any outliers?
YES
Trivial issues?
NO
Enough available metadata?
NO
Is it possible to add new metadata?
YES
Add metadata
Add information
NO
Similar data in literature, databases, etc?
YES
NO
Further research is needed
DISCARD DATA
NO
USE DATA
YES
Can outliers be corrected?
YES
Correct data
NO
YES
Study correlation outliers-metadata
Correlation with sample characteristics (population bias)?
Correlation with data origin (method bias)?
YES
NO
YES
USE DATA AS A NEW POPULATION
NO

Missing data study

- Identification of « Missingness mechanisms », *i.e.* the reasons why certain data are missing
  - **Missing At Random (MAR)**: the probability that a value is missing (« missingness ») depends on metadata present in the database (*e.g.* newer data are less likely to include Van Soest analysis)
  - **Missing Not At Random (MNAR)**: missingness depends on the value itself (*e.g.* samples with fibre analysis tend to have higher digestibility values)

Missing data study

- Extraction of a complete reference dataset (2303 samples) with no missing data for CP, CF, NDF, ADF and ADL
- Simulation of 4 incomplete sub-datasets: 2 missingness mechanisms (x 2 loss intensities (33% and 66%)



CP	CF	NDF	ADF	lignin
18,05	27,20	45,65	34,00	8,64
19,30	25,25	43,96	29,67	7,10
17,15	28,45	47,31	33,06	9,48
26,69	19,50	🟥	25,00	6,80
22,69	20,50	38,80	24,30	6,60
22,75	20,20	35,60	23,80	6,70
15,50	🟥	51,10	35,90	8,10
16,90	27,70	46,80	🟥	8,30
19,70	32,30	54,70	36,40	7,80
17,00	31,80	🟥	35,60	8,50
🟥	28,60	48,70	33,40	🟥
17,63	26,80	47,30	31,10	7,60
17,20	26,60	46,60	29,60	6,10
16,25	33,40	48,00	37,20	7,80
15,75	36,20	50,20	36,10	7,50

Missing data management methods

Deletion methods	Imputation methods
Listwise deletion All objects with a missing value in at least one variable are dropped from analysis	Mean substitution Missing data are replaced by the mean value
	Regression imputation Missing values estimated by linear regression
Pairwise deletion Only the objects with missing values in the variables involved in the analysis are dropped	Expectation-Maximization method Maximum-likelihood algorithm
	Data Augmentation method Monte Carlo algorithm (multiple imputation)

- These methods are applied to the 4 simulated incomplete datasets and the results are compared to the reference (complete) dataset:
  - Feed categorisation
  - Descriptive statistics
  - Correlations and prediction equations

Effect on feed categorisation and descriptive statistics

- Effect on feed categorisation (ANOVA)
  - Deletion methods change significantly the number of samples, masking differences between overlapping categories (hay *vs* dehydrated)
  - Imputation methods (notably Data Augmentation) can reproduce differences between hay and dehydrated at low loss intensity (33%)
- Effect on descriptive statistics
  - Deletion methods and Means substitution give significantly different descriptive statististics
  - Imputation methods tend to perform better than deletion methods, even at high loss intensity (66%)

Effect on correlations and prediction equations

- Effect on the correlation between OMD and ADF
  - Deletion methods are nearly useless in MAR situations due to the loss of ADF data. Means substitution is unsuitable too.
  - Both deletion methods and imputation methods are suitable in MNAR simulation.

Conclusion

- Feed data mining is hindered by the lack of consistent metadata and proper domain ontologies
- Outlier management
  - Univariate tests are effective to address problems allocated at the ends of the distributions
  - Multivariate tests focus on relationships between variables and can help to detect recurring error patterns
  - A heuristic approach combining formal statistical methods, *ad hoc* methods and feedback loops is recommended
- Missing data management
  - The study of missingness mechanisms may help to choose the best methods for handling missing data
  - Deletion methods are suitable with MAR data and univariate statistical analysis when the sample size is large
  - Imputation methods are useful for multivariate analysis in both MAR an MNAR contexts: they maximize information use and minimize bias

Thank you very much