Animal (2013), **7:7**, pp 1128–1136 © The Animal Consortium 2013
doi:10.1017/S1751731113000293

Data pre-processing to improve the mining of large feed databases

F. Maroto-Molina[1†], A. Gómez-Cabrera[2], J. E. Guerrero-Ginel[2], A. Garrido-Varo[2], D. Sauvant[3], G. Tran[4], V. Heuzé[4] and D. C. Pérez-Marín[2]

[1]*Servicio de Información sobre Alimentos, Universidad de Córdoba, Ctra. Nacional IV km. 396, 14014, Córdoba, Spain;* [2]*Departamento de Producción Animal, ETS Ingeniería Agronómica y de Montes, Universidad de Córdoba, Ctra. Nacional IV km. 396, 14014, Córdoba, Spain;* [3]*UMR 791 Physiologie de la nutrition et de l'alimentation, AgroParisTech, 16 rue Claude Bernard, 75231, Paris, Cedex 05, France;* [4]*Association Française de Zootechnie, AgroParisTech, 16 rue Claude Bernard, 75231, Paris, Cedex 05, France*

(Received 30 October 2012; Accepted 30 January 2013; First published online 8 March 2013)

The information stored in animal feed databases is highly variable, in terms of both provenance and quality; therefore, data pre-processing is essential to ensure reliable results. Yet, pre-processing at best tends to be unsystematic; at worst, it may even be wholly ignored. This paper sought to develop a systematic approach to the various stages involved in pre-processing to improve feed database outputs. The database used contained analytical and nutritional data on roughly 20 000 alfalfa samples. A range of techniques were examined for integrating data from different sources, for detecting duplicates and, particularly, for detecting outliers. Special attention was paid to the comparison of univariate and multivariate solutions. Major issues relating to the heterogeneous nature of data contained in this database were explored, the observed outliers were characterized and ad hoc *routines were designed for error control. Finally, a heuristic diagram was designed to systematize the various aspects involved in the detection and management of outliers and errors.*

Keywords: chemical composition, nutritive value, data integration, outlier mining

Implications

Data pre-processing receives little attention when compiling or analysing feed data. Generally, it is ignored or, at best, some standard algorithms are used without studying their implications on data mining results. This paper presents a systematic approach to data pre-processing, especially to outlier detection, highlighting its importance to ensure reliable results. We demonstrate the existence of erroneous data that slip through laboratory quality control systems, even reaching scientific papers. In addition, we point out how lack of metadata does not allow full use of data. Feed data producers, compilers and analysts should be aware of these findings to improve their future use.

Introduction

Everyday, animal feed laboratories and research centres all over the world generate countless data of various kinds from the analysis of thousands of samples of raw materials and compound feedstuffs (Gizzi and Givens, 2004). These data are generally used for a single purpose, for example, quality control or the publication of research results; they are subsequently discarded or, at best, stored in scattered and unstructured files. However, it is generally accepted that the added value obtainable from these data would more than justify the effort and cost involved in obtaining it. A number of researchers, including Tran and Lapierre (1997) and Maroto-Molina *et al.* (2011), have developed information systems intended to systematize the collection, storage and management of feed data. These systems hold hundreds of thousands of data; to make full use of them, we need expert analyses on the basis of data mining and meta-analysis tools (Sauvant *et al.*, 2010).

However, as real-world data tend to be heterogeneous, noisy and inconsistent (Han and Kamber, 2006), data pre-processing is essential to ensure reliable results. The various stages involved in pre-processing can account for up to 70% of the total time devoted to a data mining project (Pyle, 1999; Molina, 2002). Therefore, systematization of the process is one of the major challenges in current data mining research.

Numerous pre-processing techniques have been reported in the scientific literature (Kotsiantis *et al.*, 2006) often

† E-mail: g02mamof@uco.es

https://doi.org/10.1017/S1751731113000293 Published online by Cambridge University Press

accompanied by examples of their application in a wide range of fields (Piramuthu, 2006; Wu, 2009). Published papers focus mainly on outlier detection. However, very few papers address the specific use of data pre-processing methods with feed databases. In fact, few of the numerous publications dealing with data mining using feed databases (NcDowell *et al.*, 1977; Sauvant *et al.*, 2002; Tedeschi *et al.*, 2002) make any reference to the algorithms used for outlier detection, which are mentioned only tangentially and superficially. Sauvant *et al.* (2002), for example, used a 10% trimmed mean (excluding the highest 5% and the lowest 5%) when calculating the data included in their tables. Tedeschi *et al.* (2002) opted for the elimination of all data that differed by more than 2.5 standard deviations from the mean, or that inflated the variation coefficient by more than 30%. No authors in this specific field have considered the use of multivariate approaches, and no attempt has been made to type the outliers detected.

Therefore, the present study sought to evaluate a number of data pre-processing solutions to improve the data mining results of large feed databases: feed tables and prediction equations. The problems most commonly encountered are explored, and the performance of univariate and multivariate solutions is compared.

Material and methods

Data origin

The database used contained the information on alfalfa (*Medicago sativa* L.) stored in the data warehouse of the *Tables of nutritive values for farm animals in tropical and Mediterranean regions* (French acronym: TRC) (Tran *et al.*, 2010); it concerned 19 187 samples, together with their corresponding chemical and nutritional values. Data were drawn from a wide range of sources, for example, the French Association of Animal Production (French acronym: AFZ), the Feed Information Service (Spanish acronym: SIA) and the scientific literature. Information came from a total of 230 sources: 217 scientific papers and 13 databases. However, in quantitative terms, published papers accounted for only 8.2% of the samples.

Data integration

In the data warehouse of the TRC, metadata were expressed in English, French, Spanish and Portuguese, and therefore as a first step they were standardized to English. There were different forms of expression (units, descriptors, etc.) that were standardized using the system recommended by Maroto-Molina *et al.* (2008). In addition, metadata were checked to ensure that each descriptor matched a single reality (e.g. 'dry' was found to refer to 'hay' or to 'dehydrated'). ANOVA was used for this purpose. For each descriptor, main chemical components (e.g. crude protein (CP), which is available for most samples) were taken as the response and the source of data as the factor (it is assumed that each source used a single descriptor for each reality). Cases displaying a significant effect ($P < 0.05$) were checked manually. Finally, an *ad hoc* strategy for duplicate detection was designed following Hernández and Stolfo (1998). All the numeric information associated with a sample was synthesized into a key variable (KV) using the equation KV = 1 × (value first variable) + 2 × (value second variable) + … + n × (value nth variable). Samples were then ordered by KV value, checking those cases where interstitial linearity (the distance from one case to the next) equaled 0. Sum weighing coefficients in KV equation avoided false positives in samples with missing values for different variables.

Variable selection and transformation

As they store information obtained with different purposes, animal feed databases tend to be quite sparse (loosely coupled). In the alfalfa database, several variables displayed a high rate of missing values being of little value for the objective of this paper. Consequently, variables were excluded *a priori* wherever <100 data points were available. With a view to enhancing data cleansing and subsequent analysis, the dimensionality of the database was increased as much as possible. To this end, information contained in three metadata called 'original name', 'sample name' and 'observations' in the TRC data warehouse was divided into 10 fields (from 'product' to 'particle size' in Table 1).

Outlier mining

The classical definition of outlier was considered: 'an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism' (Hawkins, 1980). This does not imply an error or invalid datum (Mendenhall and Reinmuth, 1971). Several common outlier detection tests available in standard statistical software were used. Special attention was paid to the comparison of univariate and multivariate approaches.

Z-score criterion. This is probably the most widely used test. It involves establishing upper and lower limits of an acceptance interval for data. The *Z*-score was calculated for each data point as $Z\text{-score}(x_i) = (x_i - \text{mean}(X))/\text{s.d.}(X)$. All cases with $|Z\text{-score}| > 3$ were flagged as outliers (0.26% of data in a normal distribution).

Chauvenet's criterion. This test, developed by Chauvenet (1960), uses normal distribution to calculate the probability that the data point furthest from the mean actually has the assigned value, and that probability is multiplied by the number of data points; if the result is <0.5, the suspect data point is flagged as an outlier. The process is then repeated until no further outliers are detected.

Regression residuals (RegR). As the values for some attributes measured in feedstuff are interrelated, regression can be used as a mean of detecting outliers (Sauvant *et al.*, 2010). Therefore, regressions were run for the most closely related pairs of variables (minimum $r = 0.5$), and the *Z*-score criterion was applied to the studentized residuals.

https://doi.org/10.1017/S1751731113000293 Published online by Cambridge University Press

Table 1 *Fields and available data points in the database of alfalfa extracted from the data warehouse of the TRC project*

Descriptive metadata		Attributes		
Name	Available data points	Name	Unit	Available data points
Code (identifies each sample)	19 187	DM	%	17 869
Reference (source of information)	19 187	CP	%DM	18 595
Reference type (paper or database)	19 187	CF	%DM	15 840
Ecozone	16 517	ASH	%DM	9095
Country	16 517	EE	%DM	3064
Region (state, province, etc.)	3635	ST	%DM	125
Product (alfalfa or alfalfa and grass)	19 187	SG	%DM	101
Part (aerial part, stems or leaves)	19 187	NFE	%DM	2895
Process (fresh, hay, dehydrated, etc.)	19 187	NDF	%DM	3754
Grade (quality)	8817	ADF	%DM	3921
Year number	343	ADL	%DM	3116
Cut number	2808	Ca	%DM	4461
Maturity (vegetative, bloom, etc.)	604	P	%DM	4135
Age (days after last cut)	111	K	%DM	412
Variety	239	Mg	%DM	2026
Particle size (meal, pelleted, etc.)	412	Na	%DM	404
Collection day	17 235	DMdTT	%	360
Collection month	17 235	DMdpc	%	120
Collection season	17 235	DMd	%	140
Collection year	17 235	OMdTT	%	349
Animal (*in vivo* data)[a]	411	OMdpc	%	201
		OMd	%	232
		Nd	%	198
		GE	kcal/kg DM	134
		DMi	g/kg $LW^{0,75}$	128

DM = dry matter; CP = crude protein; CF = crude fiber; ASH = ash; EE = ether extract; ST = starch; SG = sugars; NFE = nitrogen-free extract; Ca = calcium; P = phosphorous; K = potassium; Mg = magnesium; Na = sodium; DMdTT = DM digestibility Tilley and Terry; DMdpc = DM digestibility pepsin-cellulase; DMd = DM digestibility *in vivo*; OMdTT = OM digestibility Tilley and Terry; OMdpc = OM digestibility pepsin-cellulase; OMd = OM digestibility *in vivo*; Nd = nitrogen digestibility *in vivo*; GE = gross energy; DMi = dry matter intake; LW = live weight.
[a]All data were obtained from ruminants.

Second principal component (PC2). Jiménez-Márquez *et al.* (2002) proposed this method. The transformation of data into principal components orthogonalizes the variance associated with data; as a result, the first component may be assumed to represent the general tendency, whereas the remaining ones represent the deviation with respect to that tendency. Principal components were calculated for the most closely related pairs of variables (minimum $r = 0.5$), and the *Z*-score criterion was applied to the second component.

Adjusted Wilks' method (aWilks). The adjustment of the Wilks (1963) method developed by Yang and Lee (1987) and implemented in MATLAB by Trujillo-Ortiz *et al.* (2006) was used. Extreme data points of a normal multivariate sample are flagged as outliers by approximating the squared Mahalanobis distance to a Snedecor f value.

Local outlier factor (LOF). This algorithm provides a local approach to outlier detection (Breunig *et al.*, 2000) on the basis of a comparison of the local density of the neighbourhood of each data point with the local density of its neighbouring MinPts (= 100 in this study). LOF does not view outlierness as a binary property; rather, it focuses on the degree of outlierness, so that LOF values close to 1 represent clear inliers, while increasing values reflect an increasing degree of outlierness. A cut-off value of 2 was selected to flag outliers with a view of comparing the results with those obtained using other algorithms.

When using the multivariate outliers aWilks and LOF, involving the simultaneous analysis of a large number of variables, the sparsity of data in the database meant that the number of variables that could be tested (i.e. those variables with no missing values) was considerably smaller. As a result, optimization of the results required staggered testing, that is, the repetition of calculations for 2, 3, 4, up to *n* variables. For the purposes of this study, tests were conducted for 4, 6 and 12 variables. Most tests (all except LOF) assume normal distribution. Only those data meeting that requirement using the Anderson and Darling (1952) test ($P < 0.10$) were used. Graphic representations (e.g. qq-graphs and scatter plots) were used to visualize the results of the various tests. Both plots and calculations were carried out using MINITAB (*Z*-score, RegR and PC2), MATLAB (Chauvenet's criterion and aWilks) and R-project (LOF).

https://doi.org/10.1017/S1751731113000293 Published online by Cambridge University Press

Table 2 *Duplicates detected in the database of alfalfa extracted from the data warehouse of the TRC project (categorized by number of available data points per sample)*

Available data points per sample	Number of samples	Duplicates	
		Number	%
1	203	9	4.43
2	1634	99	6.06
3	7809	163	2.09
4	2936	13	0.44
5	1128	10	0.89
6	1830	3	0.16
7	595	2	0.34
8	616	4	0.65
9	286	2	0.70
10	172	1	0.58
11	128	1	0.78
12	1668	6	0.36
13	63	2	3.17
14	29	0	0.00
15	78	0	0.00
16	10	0	0.00
17	2	0	0.00
Total	19 187	315	1.64

Results

Database

After data integration, selection and transformation as described above, the database used for data cleansing was shown in Table 1. The results for duplicate detection are shown in Table 2.

Outlier mining

Outliers detected using univariate criteria (*Z*-score and Chauvenet) and algorithms involving two variables (RegR and PC2) are shown in Table 3. Outliers were detected for 20 variables using *Z*-score and for 19 using Chauvenet's criterion; the mean percentages of outliers were 2.39% and 0.36%, respectively. Of the 18 variables examined using bivariate methods, RegR detected outliers in 16 and PC2 in 14. Outlier percentages were 2.33% and 1.59%, respectively. Results for aWilks and LOF tests are shown in Table 4.

Discussion

Heterogeneity

The data integration process highlighted a considerable lack of uniformity in the presentation of feed data. The following were detected:

Synonyms. Especially with regard to units and metadata descriptors, for example, 'pelleted' and 'granulated' under the metadatum 'particle size'. The standardization of this information has been widely addressed in the literature (Kalu and Fick, 1981; Moore *et al.*, 1991) but continues to be ignored in practice.

Homonyms. The same descriptors are used to denote different concepts depending on the data source. For example, within the metadatum 'maturity', the descriptor 'first cut' is used by some sources to denote the cutting carried out for weed control purposes following alfalfa planting; in other sources, this descriptor denotes the first usable harvest of alfalfa, which using the previous scale would be termed 'second cut'. The same is true for the descriptors of 'grade' for dehydrated alfalfa; some sources indicate mean protein content under this heading, whereas other take it to denote guaranteed minimum content.

Overlapping concepts. This is a recurrent issue in metadata descriptors. A good example is the use of 'bloom' and 'early bloom' under the metadatum 'maturity'. Overlapping concepts may be encountered in the data themselves, for example, to characterize *in vivo* digestibility, the concept OM digestibility (OMd) is widely used in Europe, whereas in the United States the different components of the Weende analysis tend to be characterized separately. Integration of all this information in a single scale requires the use of the algorithm $OMd = (CP \times CPd + CF \times CFd + EE \times EEd + NFE \times NFEd) \times 100/(100–ASH)$. In any case, mining all available metadata and data always entails a certain loss of information, as data always have to be integrated using the less-specific scale. Sometimes, even though data refer to an equivalent concept, they cannot be integrated into a single scale. A good example is provided by the metadata 'maturity' and 'age', both of which denote the physiological development of the plant, but using incompatible descriptors; a given plant age does not correspond to a single stage of maturity, as many other factors (genetics, soil–climate combinations, etc.) influence plant development (Mueller and Teuber, 2007).

Ambiguous concepts. The meaning of some descriptors may be difficult to interpret, for example, the descriptor 'mature' within the metadatum 'maturity'. Abreu *et al.* (2000) use this term to denote the physiological point at which pods are developed and seeds are milky or soft; yet, as alfalfa is not generally harvested at such an advanced stage, 'mature' could be used to denote 'non-vegetative', that is, any stage from early bloom onwards.

Duplicates

In samples containing a large number of data points, the exact coincidence of all data points clearly indicates repetition; however, where a sample contains few data points, this may be no more than a chance coincidence, given the size of the database. Although it is not possible to establish a cut-off value for the number of data points per sample, the use of KV and interstitial linearity reduces considerably the number of potential duplicates to manually check.

The duplicates identified did not correspond, as might be expected, to bibliographical information published on more

https://doi.org/10.1017/S1751731113000293 Published online by Cambridge University Press

Table 3 *Outliers detected by univariate and bivariate methods in the database of alfalfa extracted from the data warehouse of the TRC project*

		Data flagged as outliers by univariate methods		Data flagged as outliers by bivariate methods		
Attributes	Available data points	*Z*-score	Chauvenet	Revised data points[b]	RegR	PC2
DM	17 869	1304	0	123	0	0
CP	18 595	371	124	16 873	421	271
CF	15 840	134	29	15 718	359	239
ASH	9095	106	69	4401	74	48
EE	3064	29	19	165	4	0
NFE	2895	29	6	19	0	0
ST[a]	125	–	–	–	–	–
SG[a]	101	–	–	–	–	–
NDF	3754	27	2	3680	42	34
ADF	3921	33	4	3921	172	134
ADL	3116	22	11	2919	79	67
Ca	4461	31	4	4401	74	48
P[a]	4135	62	36	–	–	–
K[a]	412	3	2	–	–	–
Mg[a]	2026	1	1	–	–	–
Na	404	5	4	165	4	0
DMdTT	360	4	4	325	4	4
DMdpc[a]	120	–	–	119	4	2
DMd	140	0	0	133	1	1
OMdTT	349	5	3	337	9	4
OMdpc	201	0	0	198	4	3
OMd	232	3	3	215	4	1
Nd	198	2	2	165	3	3
GE[a]	134	5	5	–	–	–
DMi[a]	128	4	3	–	–	–

DM = dry matter; CP = crude protein; CF = crude fiber; ASH = ash; EE = ether extract; ST = starch; SG = sugars; NFE = nitrogen-free extract; Ca = calcium; P = phosphorous; K = potassium; Mg = magnesium; Na = sodium; DMdTT = DM digestibility Tilley and Terry; DMdpc = DM digestibility pepsin-cellulase; DMd = DM digestibility *in vivo*; OMdTT = OM digestibility Tilley and Terry; OMdpc = OM digestibility pepsin-cellulase; OMd = OM digestibility *in vivo*; Nd = nitrogen digestibility *in vivo*; GE = gross energy; DMi = dry matter intake.
[a]These variables were not checked for outliers because they either displayed statistically non-normal probability distributions or were not included in any pair of variables using established correlation criteria (or both).
[b]The number of data points revised by bivariate methods for each variable depends on the number of common points and the correlation (must be higher than 0.5) of each variable with the rest.

Table 4 *Outliers detected by multivariate methods in the database of alfalfa extracted from the data warehouse of the TRC project (as a function of number of variables considered by the algorithm)*

Test	Number of variables[a]	Revised samples[b]	Samples flagged as outliers
aWilks	4	7788	45
	6	2770	27
	12	1543	47
LOF	4	7788	163
	6	2770	122
	12	1543	83

aWilks = adjusted Wilks; LOF = local outlier factor.
[a]The high percentage of missing values in the alfalfa database meant that the number of variables that could be tested with a multivariate method decreased as the number of variables considered by the method increased. For the purposes of this paper, tests were conducted considering 4, 6 and 12 variables (of a total of 25).
[b]Of a total of 19 187 samples of alfalfa.

than one occasion or included in more than one database, but rather to laboratory data entered on more than one occasion in databases. Some databases are built up using input from laboratories, few of which operate traceability systems that would prevent the same data being resent to database managers.

Outlier mining

In the case of univariate methods, the *Z*-score criterion detected more outliers than the Chauvenet's criterion, although when examined variable-by-variable, differences between the two techniques diminished as the number of data points decreased. Differences in the number of outliers

https://doi.org/10.1017/S1751731113000293 Published online by Cambridge University Press

detected for the variable dry matter (DM) were particularly large. Graphical analysis showed that most *Z*-score outliers corresponded to green alfalfa and silage samples, which were less numerous than hay and dehydrated alfalfa samples, the latter displaying much higher DM values (70% to 90% *v.* 10% to 40%). This highlights the need to take descriptive metadata into account during outlier detection; He *et al.* (2004) named this class outlier mining. Regarding multivariate methods, more outliers were detected by RegR than by PC2, and more by LOF than by aWilks.

When selecting a test, other aspects but the number of outliers must be considered, including those relating to algorithm implementation. RegR entails twice as many calculations as PC2, as residuals are not symmetrical, that is, they differ depending on whether each variable is a pair acting as a dependent or an independent variable. On the other hand, calculations using aWM and LOF are simpler than those involved in RegR and PC2: a simple algorithm for the whole database *v.* different algorithms for each pair of variables. However, the elimination of an outlier detected using a bivariate method entails the loss of two data points, whereas in a test with 12 variables 12 data points will be lost. In addition, as indicated earlier, in this type of test involving the simultaneous analysis of a large number of variables, the sparsity of data in the database meant that the number of variables that could be tested was much smaller. Using bivariate methods, a high percentage of data points were tested for those variables with many common points and a strong correlation with the rest (e.g. ADF), but the number of data points tested for other variables (e.g. DM) was considerably reduced (Table 3). Using aWilks and LOF tests for 12 variables (of a total of 25) meant that barely 8% of samples could be tested (Table 4).

In all cases, the percentage of detected outliers was lower than that reported by other authors (Wang *et al.*, 1995; Müller and Freytag, 2003), usually between 5% and 10%. This is because some of the data in the database had already been cleaned using both formal and non-formal methods. Roughly, 75% of samples were drawn from the AFZ database, for which 10% of data supplied by laboratories had been removed (Tran and Lapierre, 1997) before entering the database. Therefore, the outliers detected were additional to that 10%.

Consistency analysis

Attention must be paid to the consistency of those data points flagged as outliers by the various methods tested. Table 5 illustrates findings for the subgroup of 1543 samples comprising 12 variables with no missing data. All the outliers detected by the Chauvenet criterion were also detected by the *Z*-score method; the same is true for PC2 and RegR. Thus, the essential difference between these methods is one of rigor. However, differences between aWilks and LOF were much greater. Although results for aWilks were consistent with those obtained using bivariate methods, LOF detected most of the outliers identified using the other techniques, as well as 14 not detected by them.

Table 5 *Number of samples with common outliers detected by different methods (univariate and multivariate) in the database of alfalfa extracted from the data warehouse of the TRC project*

Criteria	*Z*-score[a]	Chauvenet	RegR	PC2	aWilks	LOF
Z-score[a]	157					
Chauvenet	73	73				
RegR	94	57	140			
PC2	72	46	95	99		
aWilks	43	33	40	37	47	
LOF	61	30	44	38	24	83

RegR = regression residuals; PC2 = second principal component; aWilks = adjusted Wilks; LOF = local outlier factor.
[a]Outlier *Z*-score using class-outlier mining with metadata 'process'.
The diagonal represents the number of samples in which one or more outliers were detected by each method. The other cells show the number of these samples in which the other tests also detected outliers.

Qualitative characterization of outliers

Analysis of the characteristics of the outliers flagged by the various methods (often involving consultation of the original source) yielded the following classification:

Transcription errors. These are associated with the manual handling of data, the most evident error being the transposition of a decimal point or the use of inappropriate units. Descriptive metadata may also be affected.

Interpretation errors. A good example is related to nutritional attributes, for example, digestibility: often, particularly, if data are collected by inexpert staff, *in vivo* information is confused with *in vitro* measurements, or even with data estimated using prediction equations. As most of these errors are not detected by standard outlier tests, *ad hoc* analysis is required. In this respect, scatter plots can prove to be very valuable (e.g. in Supplementary Figure S1: data from Morocco are calculated, not *in vivo*), but the systematization of the process requires comparison of general and intra-source correlations.

Procedure issues. The problem sometimes lies in the sampling technique used. Results for ASH, for example, frequently yield a long-tailed distribution in which the highest values are found for soil-contaminated samples. Sometimes, the laboratory procedure itself generates anomalous data. Hatfield and Fukushima (2005) listed a range of factors likely to prompt an overestimation of lignin values. Similarly, forage fat content is frequently overestimated because of the inclusion in the ether extract (EE) residue of non-saponifiable substances such as waxes, cutins and chlorophyll (Palmquist and Jenkins, 2003).

Uncommon values. Not all outliers are errors; sometimes, they merely indicate exceptional data. Outliers of this kind generally appear at the edges of probability distributions. Alfalfa samples (especially hays) with a low leaf-to-stem ratio and very mature samples will appear at the low end of the CP distribution, whereas very young samples, urea-treated silage

https://doi.org/10.1017/S1751731113000293 Published online by Cambridge University Press

Table 6 Ad hoc *error-detection filters designed on the basis of the analysis of the features of outliers flagged by standard methods (univariate and multivariate) in the database containing the information on alfalfa stored in the data warehouse of the TRC project plus expert knowledge of the relationships between variables*

Ad hoc error-detection filters	Mandatory	Number of errors detected
Mathematical		
Any data must be higher than zero	Always	3
Data expressed as % must be lower than 100	Always	0
Data expressed as g/kg must be lower than 1000	Always	0
Nutritional		
NDF must be higher than ADF	Most times[a]	9
ADF must be higher than ADL	Most times[a]	0
NDF must be higher than ADL	Always	0
ADF must be higher than CF	Many times	59
NDF must be higher than CF	Most times	1
ASH must be higher than the sum of minerals	Always	1
OMd *in vivo* must be higher than DMd *in vivo*[b]	Always	9
Difference between OMd *in vitro* and DMd *in vitro* must be within the range established by: DMd × (ASH/OM) and DMd × (ASH/OM)–100(ASH/OM)[c]	Always	147

CF = crude fiber; OMd = OM digestibility; DMd = DM digestibility.
[a]It might be thought that these filters should be set to 'always', as these differences represent hemicellulose and cellulose content, respectively, which can never be negative. However, this flexibility seeks to reflect certain constraints affecting Van Soest fractionation: starch retention (solved in part by the addition of amylases), hemicellulose solubilization in NDF residue and pectin/tannin retention in ADF residue. As a result, ADF may be greater than NDF in samples with high pectin and/or tannin content.
[b]The difference between the two measurements reflects the *in vivo* digestibility of ash, which should be very low or non-existent (it may even be negative, because of the mineral supply from endogenous secretions from the digestive tract or to the presence of a mineral source other than the feedstuff evaluated, e.g. a mineral–vitamin corrector).
[c]The range is determined by the fact that, for *in vitro* digestibility, the digestibility of ash may vary between 0 and 100.

and high-quality dehydrated alfalfa will appear at the upper end. The anomaly sometimes resides not in the sample itself but in the experimental conditions. In the alfalfa database, for example, low digestibility data points were recorded owing to large intake, and thus faster passage through the digestive system.

Error discrimination

The standard tests evaluated cannot distinguish errors from outliers. However, analysis of the features of the outliers flagged using these tests enabled certain recurring error patterns to be identified. This information plus expert knowledge can be used to design *ad hoc* filters for the detection and elimination of certain errors. The filters constructed for the studied database are shown in Table 6, indicating the number of errors detected with each filter. These filters represent an improvement with regard to standard outlier tests, as they detect only errors, many of which escape the tests in question: an example is given in Figure 1, regarding the relationship between *in vitro* OM and DM digestibilities.

Effects on database outputs

Feed tables normally include mean values calculated from the information stored in feed databases. The deletion of errors detected by standard tests and *ad hoc* filters has little effect on mean values, even in those variables having a high rate of errors. For example, for OM digestibility Tilley and Terry (OMdTT) (25% of errors), the mean value calculated with all data was 62.05%, whereas the mean value for

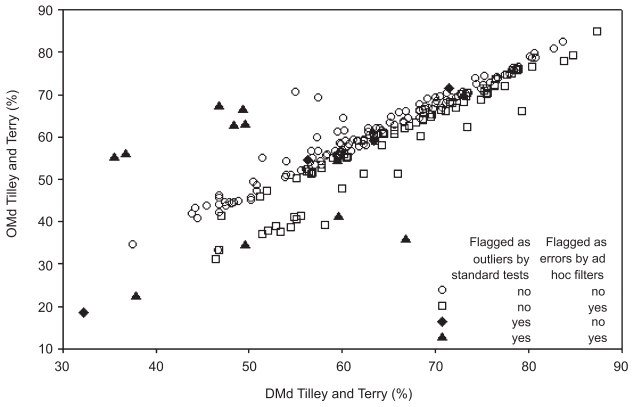


Figure 1 Comparison of data points flagged as outliers using standard tests and as errors using *ad hoc* filters. Many data points are not flagged as outliers by standard tests, but they are flagged as errors by *ad hoc* filters. Some of these data have a strong influence in the relationship between OMdTT and DMdTT.

cleaned data was 63.33%. However, deletion of errors may have an important effect on range and s.d. values, especially because of the elimination of extreme data (detected by univariate tests). Prediction equations are another important output of feed databases. In this case, erroneous data may influence the relationships between variables as was shown in Figure 1. Selection of outlier tests should be made according to data use. If data are used only for description purposes (feed tables), univariate outlier tests may be enough. However, if data are used to build models

https://doi.org/10.1017/S1751731113000293 Published online by Cambridge University Press

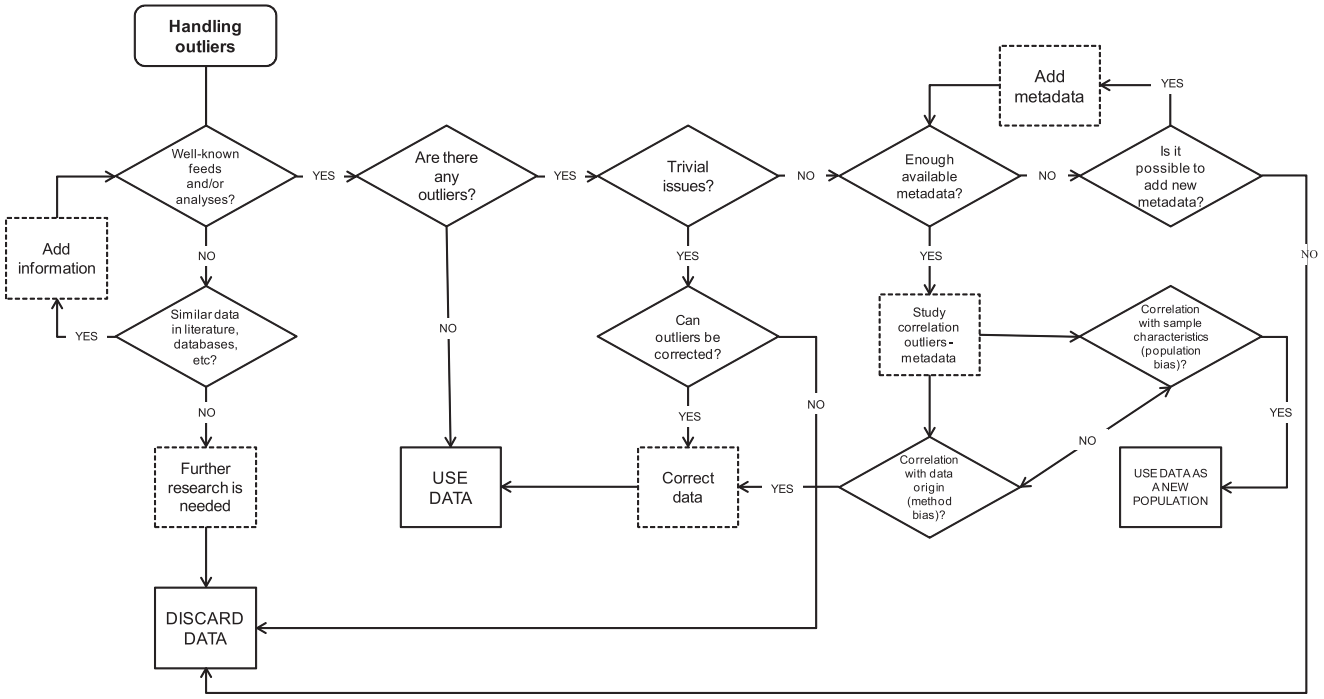


Figure 2 Heuristic diagram for the management of outliers and errors in feed databases.

(prediction equations), they must be tested by multivariate and *ad hoc* approaches.

Heuristic diagram

Using the information described in this paper, a heuristic diagram was designed for outlier detection and management (Figure 2). Owing to the space limitations, a single block has been used to represent all the outlier finding tests described in the present study: univariate, multivariate and *ad hoc*. Outlier detection is not a linear process but algorithms should be run multiple times and their results compared with each other.

Conclusions

The heterogeneous descriptions of feed data and metadata limits their potential for data mining and meta-analysis purposes, as sometimes it is not possible to integrate all the available data and when possible, it normally implies loss of information.

A wide range of outliers and errors tend to be detected in animal feed databases. Outlier interpretation is often hindered by the lack of metadata, and thus requires consultation of the original data sources, which are not always available. Thus, metadating is of particular importance for the interpretation of exceptional data. In this field, there is a need of comprehensive and well-defined metadata, as well as proper domain ontologies.

The standard outlier tests are unable, used alone, to detect all kinds of outliers: univariate tests are more effective when addressing problems allocated at either end of the probability distribution, whereas multivariate tests focus on the relationships between variables. In addition, the supplementary use of multivariate methods proved effective for locating recurring error patterns, which can be used as a basis for designing *ad hoc* error-detection filters.

The choice of an outlier test is dictated by the purpose of subsequent analysis. When selecting a test, issues such as rigor and process implementation should be taken into account. Feed data pre-processing cannot be based on a single method, but instead it should consist in a heuristic approach relying on a combination of general-purpose formal statistical methods, experience-based *ad hoc* algorithms and feedback loops.

Acknowledgements

The help of Alejandro Argamentería-Gutierrez, PhD (Regional Service for Agri-food Research and Development, Asturias) in the interpretation of the characteristics of laboratory measurements is gratefully acknowledged. The authors are also grateful to the Spanish Ministry of Science and Innovation for a technical assistant grant to the Feed Information Service.

Supplementary materials

For supplementary material referred to in this article, please visit http://dx.doi.org/10.1017/S1751731113000293

References

Abreu JM, Bruno-Soares AM and Calouro F 2000. Intake and nutritive value of Mediterranean forages and diets: 20 years of experimental data. ISA Press, Lisbon, Portugal.

Anderson TW and Darling DA 1952. Asymptotic theory of certain 'goodness-of-fit' criteria based on stochastic processes. Annals of Mathematical Statistics 23, 193–212.

https://doi.org/10.1017/S1751731113000293 Published online by Cambridge University Press

Breunig MM, Kriegel HP, Ng RT and Sander J 2000. LOF: identifying density-based local outliers. Retrieved May 15, 2012, from www.dbs.informatik.uni-muenchen.de/Publikationen/Papers/LOF.pdf.

Chauvenet W 1960. A manual of spherical and practical astronomy. Dover Publications, New York, USA.

Gizzi G and Givens DI 2004. Variability in feed composition and its impact on animal production. Retrieved February 25, 2011, from www.fao.org/docrep/Article/Agrippa/X9500E03.HTM.

Han J and Kamber M 2006. Data mining: concepts and techniques. Elsevier, San Francisco, USA.

Hatfield R and Fukushima RS 2005. Can lignin be accurately measured? Crop Science 45, 832–839.

Hawkins DM 1980. Identification of outliers. Chapman and Hall, London, UK.

He Z, Xu X, Huang JZ and Deng S 2004. Mining class outliers: concepts, algorithms and applications in CRM. Expert System with Applications 27, 681–697.

Hernández MA and Stolfo SJ 1998. Real-world data is dirty: data cleansing and the merge/purge problem. Data Mining and Knowledge Discovery 2, 9–37.

Jiménez-Márquez SA, Lacroix C and Thibault J 2002. Statistical data validation methods for large cheese plant database. Journal of Dairy Science 85, 2081–2097.

Kalu BA and Fick GW 1981. Quantifying morphological development of alfalfa for studies of herbage quality. Crop Science 21, 267–271.

Kotsiantis SB, Kanellopoulos D and Pintelas P 2006. Data pre-processing for supervised learning. International Journal of Computer Science 1, 111–117.

Maroto-Molina F, Gómez-Cabrera A, Guerrero-Ginel JE and Garrido-Varo A 2008. Propuesta para la homogenización de la información sobre alimentos: aplicación a la base de datos Pastos Españoles (SEEP). Pastos 38, 141–184.

Maroto-Molina F, Gómez-Cabrera A, Guerrero-Ginel JE, Garrido-Varo A and Pérez-Marín DC 2011. Building a metadata framework for sharing feed information in Spain. Journal of Animal Science 89, 882–888.

Mendenhall W and Reinmuth JE 1971. Statistics for management and economics. Duxury Press, Belmont, USA.

Molina LC 2002. Data mining: torturando los datos hasta que confiesen. Retrieved March 12, 2012, from www.uoc.edu/web/esp/art/uoc/molina1102/molina1102.html.

Moore KJ, Moser LE, Vogel KP, Waller SS, Johnson BE and Pedersen JF 1991. Describing and quantifying growth stages of perennial forage grasses. Agronomy Journal 83, 1073–1077.

Mueller SC and Teuber LR 2007. Alfalfa growth and development. Retrieved July 4, 2011, from alfalfa.ucdavis.edu/IrrigatedAlfalfa/pdfs/UCAlfalfa8289GrowthDev_free.pdf.

Müller H and Freytag J 2003. Problems, methods and challenges in comprehensive data cleansing. Retrieved September 30, 2012, from http://www.dbis.informatik.hu-berlin.de/fileadmin/research/papers/techreports/2003-hub_ib_164-mueller.pdf.

NcDowell LR, Conrad JH, Thomas JE, Harris LE and Fick KR 1977. Nutritional composition of Latin American forages. Tropical Animal Production 2, 273–279.

Palmquist DL and Jenkins TC 2003. Challenges with fats and fatty acid methods. Journal of Animal Science 81, 3250–3254.

Piramuthu S 2006. On pre-processing data for financial credit risk evaluation. Expert System with Applications 30, 489–497.

Pyle D 1999. Data preparation for data mining. Morgan Kaufmann, San Francisco, USA.

Sauvant D, Pérez JM and Tran G 2002. Tables de composition et de valeur nutritive des matieres premieres destinees aux animaux d'elevage. INRA Editions, Paris, France.

Sauvant D, Tran G, Heuzé V, Bastianelli D and Archimède H 2010. Data engineering for creating feed tables and animal models in the tropical context. Advances in Animal Biosciences 1, 438–439.

Tedeschi LO, Fox DG, Pell AN, Duarte DP and Boin C 2002. Development and evaluation of a tropical feed library for the Cornell net carbohydrate and protein system model. Scientia Agricola 59, 1–18.

Tran G and Lapierre O 1997. The French feed database: a national network for collecting and disseminating data about feedstuff composition and nutritive value. In First European Conference for Information Technology in Agriculture (ed. H Kure, I Thysen and AR Kristensen), pp. 105–108. Copenhagen, Denmark.

Tran G, Heuzé V, Bastianelli D, Archimède H and Sauvant D 2010. Tables of nutritive value for farm animals in tropical and Mediterranean regions: an important asset for improving the use of local feed resources. Advances in Animal Biosciences 1, 468–469.

Trujillo-Ortiz A, Hernández-Walls R, Castro-Pérez A and Barba-Rojo K 2006. MOUTLIER: Detection of outlier in multivariate sample test, a MATLAB file. Retrieved June 15, 2012, from www.mathworks.com/matlabcentral/fileexchange/122522.

Wang RY, Reddy MP and Kon HB 1995. Towards quality data: an attribute based approach. Decision Support Systems 13, 349–372.

Wilks SS 1963. Multivariate statistical outliers. Indian Journal of Statistics 25, 407–426.

Wu Z 2009. A review of statistical methods for pre-processing oligonucleotide arrays. Statistical Methods in Medical Research 18, 533–541.

Yang SS and Lee Y 1987. Identification of a multivariate outlier. Paper presented at the Annual Meeting of the American Statistical Association, August 1987, San Francisco, USA.

https://doi.org/10.1017/S1751731113000293 Published online by Cambridge University Press