Missing value imputation and data cleaning in untargeted food chemical safety assessment by lc-hrms

Grégoire Delaporte, Mathieu Cladière, Valérie Camel

Missing value imputation and data cleaning in untargeted food chemical safety assessment by LC-HRMS

Grégoire Delaporte, Mathieu Cladière*, Valérie Camel

UMR Ingénierie Procédés Aliments, AgroParisTech, Inra, Université Paris-Saclay, 91300 Massy, France

** Corresponding author: AgroParisTech, 16 rue Claude Bernard, F-75005, Paris, France*

Phone: +33 1 44 08 37 01 – email: mathieu.cladiere@agroparistech.fr

Abstract:

Untargeted food safety assessment by the use of LC-HRMS instrumentation combined to chemometric tools is a rather new field. As a consequence there is a lack of methodological assessment of the different steps of the data treatment workflow. Thus, we propose a comparison of different methods applied to two major steps of data matrix pretreatment, namely missing value imputation and ion selection. To that end, a missing value classification method has been proposed for the first time for MS data. Several metrics have also been proposed to assess pretreatment step performance as well as to investigate global untargeted approach efficiency for all method combinations considered. Different contaminants were considered as "tracers" to address their detection rates. Pretreatment methods were applied here on two data sets, aiming at illustrating either a simple contamination case to detect or a more complicated application. The data sets used in this study were from the EML-EBI Metabolights data exchange platform (MTBLS752 and MTBLS754), offering other research groups the opportunity to develop and compare their own data treatment strategies with the combinations discussed in this work.

Keywords: LC-ToF-MS; food contaminants; non-targeted; variable selection; filtration

1. INTRODUCTION

Due to the complexification of food production chain and market, and the growing demand of consumers for safer food products, the development of new untargeted analytical strategies for food chemical safety assessment emerged over the last years [1–5]. To that end, high resolution hyphenated instruments such as UHPLC-HRMS combined with chemometric methods were identified as highly promising tools, since they had already been applied to detect and characterize unknown or unexpected compounds in metabolomics studies [6,7]. However, their adaptation to food chemical safety assessment raises many challenges due to the complexity of food samples and the trace levels of contaminants [8–10].

Untargeted analyzes do generate highly complex signal mixtures, often composed of several thousand ions after peak extraction and alignment [10]. In chemical food safety applications, the user is often interested in only few dozens of those signals, related to chemical contaminants or residues. Those signals of interest are most of the time of much lower intensity compared with other signals present (especially those related to food constituents), meaning that strong data filtration approaches [2,5,10] coupled to powerful data exploration strategies [3,11] and multivariate methods [1,5,8] must be set up to detect potential contaminants. Inappropriate filtration methods may lead to either false negative results (compounds of interest are removed from the data matrix) or unusable data matrix (too much interfering compounds remain in the data matrix). In that view, strategies based on univariate statistics coupled to the use of a fixed fold change (FC) threshold have been proposed [2]. Another approach of data filtration of metabolomics-like LC-MS data sets has been proposed recently [12], based on the calculation of a minimum relevant FC (FC_{min}) from which a signal difference can be considered as significant for each peak. Thus, the comparison between this new approach and the strategy based on the t-test / fixed FC combination should bring interesting outcomes in untargeted chemical food safety assessment studies.

Moreover, despite the performances of analytical methods and peak extraction algorithms, missing values are frequently found in final data matrices [13]; they are of great concern in untargeted approaches since they may represent around 20% of all values in MS-based data sets [14]. Missing

values are generally classified into three categories [15]: (i) Missing Completely at Random (MCAR) that occur randomly and independently to other variables, (ii) Missing at Random (MAR) that occur randomly but for which the probability of missing is influenced by other variables, (iii) Missing Not at Random (MNAR) for peaks below the detection capability of the instrument or below minimum criteria of the peak extraction algorithm. In MS-based data sets, MCAR and MAR cannot be distinguished since they are due to errors in the measurement or peak extraction process [13,15]; therefore, they will be considered as a unique MAR category in this study.

Bad handling of missing values is known to lead to poor outcome of the data process [13,14,16]. Comparison of missing value imputation methods has been recently reported for LC-MS metabolomics data sets [13,14,16]: imputing a single value (for example zero or the median of measurements) to all missing values gave poor outcomes; another approach is to use data analysis tools and multivariate methods to predict missing values. Last but not least, missing values can be imputed by a forced peak integration of the raw data: this strategy is implemented within the XCMS R package ("xcms.fillPeaks" module) [17]. Compared with previously described methods, the values provided by this latter approach should be closer to reality; however, with HRMS technologies, missing peaks may generate a total absence a signal (i.e. a flat baseline) and further a high amount of zero values in the raw data set, with subsequent numerous MNAR values. While efficiency of single value and multivariate imputation methods have already been discussed for metabolomics studies [13,14,16], xcms.fillPeaks has never been compared to the other approaches. Also, several works suggest that MNAR and MAR should be implemented by different methods for LC-MS data sets [13,15], which is not the case for reported studies [14,16].

Two approaches can be reported for missing values study. The first, used by Wei et al [13], consists in using a complete data set in which missing values are artificially generated, their distribution being controlled. This offers the advantage of easily making a fine assessment of missing value imputation methods, but the distribution of missing values in the data set may be different than for "native" ones. The second approach, used by Di Guida et al [16], relies on the use of benchmark data sets, on which several data treatment processes featuring various missing value imputation methods are applied. In this

case, the performance assessment is more difficult and rely on global performance index of the approach (e.g. detection rate) or intermediate metrics. However, this latter approach enables the implementation of the methods in "real-life" cases and should give a more realistic, even though less fine, overview of the method performances. So, our work is based on real data sets that contain native missing values.

As spotted by Di Guida et. al. [16] for metabolomics studies, an assessment of the whole workflow and of the influence of each step on its outcome is complementary to the study of the tools themselves to propose guidelines, since the quality of each step is highly linked to the one of the previous. Nowadays, even though the global workflow for untargeted food safety assessment using a metabolomics-like approach seems to be more or less established [8,10], there is a lack of vision on the influence of the different tools used for each step on the performance of the whole process. So, this paper aims at giving an overview of the influence of two important steps in the data treatment, imputation of missing values and filtration of data matrix. For the first time to the best of our knowledge, a missing value classification method was proposed for MS data. This classification method was used to set up missing value imputation approaches by combining existing imputation methods (namely "mean-LOD" and "SVD-QRILC"), which were compared to the fillPeaks tool of the XCMS package (which is a classical reference missing value imputation method for LC-MS data). For data filtration, a method commonly used in untargeted food safety studies based on t-test and fold change calculation with fixed filtration thresholds was compared with a rather new one coming from the field of metabolomics, based on the calculation of a minimum relevant fold change for each ion [12]. Resulting data treatment processes were applied on different UHPLC-HRMS data sets related to untargeted food chemical safety assessment and their respective performances presented and discussed.

2. MATERIAL AND METHODS

The influence of three missing value imputation and two filtration methods (leading to six different combinations) has been assessed as a part of an existing data treatment workflow developed for untargeted food contaminants detection [5].

2.1 DATA SETS

Unlike metabolomics studies, there is currently no data set on untargeted food contaminants detection available online excepted two data sets recently deposited by our team on the EMBL-EBI MetaboLights database (DOI: 10.1093/nar/gks1004. PubMed PMID: 23109552 [19]) with the identifiers MTBLS752 (data set #1 https://www.ebi.ac.uk/metabolights/MTBLS752) and MTBLS754 (data set #2 https://www.ebi.ac.uk/metabolights/MTBLS754) [5]. Each data set is composed of two sub-sets, one for each ionization mode. These two house data sets were selected for the present study since the lack of others available online makes impossible the discussion on other data sets.

Both data sets are based on tea samples. Green tea leaves (*camellia sinensis*) samples from two brands were bought from local stores: green tea n°1 is an organic Bancha tea from Japan and green tea n°2 is a conventional farming tea from China. Tea samples were spiked at several levels (from 10 to 100 µg/kg) (3 preparation replicates per level) with two mixes of contaminants (plus a mix of isotopically labelled molecules to check the quality of the analysis) (see Table 1). They were further analyzed using a generic sample treatment (direct solvent extraction and concentration) followed by broad range UHPLC-HRMS method [18] (Waters H-Class UPLC system coupled with a Waters Xevo G2-S ToF mass spectrometer equipped with and electrospray ion source in positive and negative ion centroid mode, *m/z* range from 60 to 800) described in supplementary materials. Each sample preparation replicate was injected three or four times (depending on the data set), data files originating from samples of same brands and same spiking levels being called “group” (n=9 or 12; 3 sample replicates analyzed 3 - 4 times each). Injection orders were randomized, and each data set also includes blank (solvent) injections as well as quality control samples (QC, pooled extracts) injected regularly (every 10 or 15 injections depending of the data set).

Each data set presents a different challenge. The contamination is expected to be easy to detect in data set #1, due to the presence of numerous molecules in the spiking mix. For this data set, the main question will be on the detection rate obtained with each method combination. In data set #2, to distinguish

between the variability due to the spiking and the one caused by the sample is likely to be the main challenge.

An additional data set has also been used to discuss the behavior of fillPeaks algorithm on data exhibiting flat baselines for some ions. It is also related to tea samples spiked with several food contaminants at low levels, but in that case analyzes were conducted on a LC-Orbitrap platform. Only data files acquired in positive ionization mode were used. Experimental details on this data set, as well as raw data files can be found on Metabolights data repository with the identifier MTBLS771 (https://www.ebi.ac.uk/metabolights/MTBLS771).

2.2 DATA TREATMENT WORKFLOW

The data treatment workflow is described in Figure 1. It can be divided into four main steps: A - building the data matrix, B – preparation and pretreatment of the built matrix (i.e. handling of missing values and ions filtration), C - Scaling and normalization, D - multivariate analysis and suspect ions annotation. Data files were firstly converted to mzXML format using Proteowizard [20], and then uploaded on the Galaxy/Workflow4Metabolomics (W4M) platform [21] where the data matrix was built using the CentWave algorithm of the XCMS package [17,22] (a full list of XCMS parameters can be found in Supplementary material Table S.2). The data matrix is composed of the peak areas for the different replicates for every ion (i.e. variable) characterized by its retention time (RT) and *m/z*. At this point (between step A and B in Figure 1), metrics on missing values (detailed in 2.3) were calculated. Missing values (MV) were then imputed either on W4M, RStudio (Version 1.1.383, R version 3.4.1) or in Matlab (Matlab 7.5.0, 2007b, The MathWorks) depending on the imputation method used.

All steps further were done in Matlab. After the filtration, the data matrix undergoes a normalization and scaling step (step C in Figure 1: log, Pareto and Probabilistic Quotient Regression –PQN– were applied). Finally a multivariate method (Independent Component Analysis, ICA [5,23]) was implemented to highlight a potential separation of groups. Thanks to ICA, group separations could be linked to corresponding ions which were then automatically annotated using a data mining method to detect isotopic patterns [3,5] followed by a broad range in-house built database search. At the end, the

annotation of discriminating ions was manually curated and the found contaminants compared with the ones spiked (called "tracers"), enabling a detection rate of our "tracers" to be estimated.

2.3 MISSING VALUE CLASSIFICATION AND METRICS

In most MS-based metabolomics studies [14, 16], missing values are all imputed using the same method, either simple (e.g. all missing values are imputed with zero or the median of all measurements), or more complex (e.g. missing values are predicted using multivariate statistical methods). Even though the multiplicity of nature of missing values is well known by statisticians for a long time, its implication in MS data sets has been only raised in 2016 in the field of proteomics [15], and even more recently in the field of metabolomics [13]. However, until now, there is no methodology to classify missing values in MS data sets. Yet, as spotted by Wei et al. [13], it is important to differentiate missing values depending on their nature. Thus, we used injection replicates for each sample preparation replicate to determine the nature of the missing values, according to the following (for each ion and sample):

- Missing at Random (MAR) when, for one sample preparation replicate, there is only one missing value among the 3 or 4 injection replicates;
- Missing not at random (MNAR) when, for one sample preparation replicate, there is more than one missing value among the injection replicates (i.e. value near or below the detection capability of the overall method).

The proposed classification methodology for missing values is represented in Figure 2. Although these classification criteria can surely be improved and discussed, they have the advantage to be consistent with the performance of the instrumentation used in terms of stability and repeatability, and also to be easily applied to large data sets. Thanks to this methodology, it is now easy to pick the best method for each category (MNAR or MAR). Another advantage is the easy combination with any existing missing value imputation methods, simple or complex, including new ones.

This classification was done after the peak extraction and alignment step (step A in Figure 1), and several metrics were then calculated on each data set: global, group-wise and category-wise missing value rates were calculated. To assess the distribution of missing values in data sets, Pearson correlation coefficients

between the frequency of missing values and *m/z*, RT or mean areas were calculated. Missing value frequencies were also plotted against *m/z*, RT or mean areas to assess any potential trend which could not be detected only by correlation coefficients [14,16].

2.4 MISSING VALUE IMPUTATION STRATEGIES

In this work, three different methods were picked for missing value (MV) imputation: one imputes all missing values at once by forced integration of the raw chromatogram while two impute separately MAR and MNAR.

The first method replaces all MV by values estimated upon signal integration in the RT window of the missing peak in the raw data files. This was automatically performed using in-line implementation of the fillPeaks method on W4M platform.

The second method (named "mean-LOD") imputes MV separately with simple strategies. MAR are imputed by the mean of the non-missing replicates of the concerned ion and a noise component, with a random relative standard deviation (RSD) between -20% and +20% around the mean value (approximately corresponding to the observed RSD on reliable peaks on pool samples in the data set), is added to limit its influence on the following steps of the process. MNAR are imputed by the limit of detection (LOD) of the instrumental method, calculated here as the mean of the 3% lowest non-missing values [24], while adding the same noise component as for MAR.

The third method (named "SVD-QRILC") imputes MAR and MNAR separately with methods based on statistical tools, respectively singular value decomposition (SVD) for MAR [13] and quantile regression imputation of left-censored data (QRILC) for MNAR. For SVD method, MV are firstly initialized to 0 and then estimated through an iteratively application of an eigen-values decomposition: here, the R wrapper based on the function "pcaMethod" has been used [25]. Another method (Random Forest, RF), possibly better than SVD for MAR [13], has been assessed but either the size of our data sets (more than 20,000 ions for around 40 samples) or the MV distribution was such that computation time was too high (no convergence was observed after 24 h of computing against a convergence achieved in a few dozen of seconds for SVD) for its application here. Hence, we were not able to compare the results of this

algorithm based on a learning method to the other proposed methods due to insufficient computing power despite the use of a computational server. So, the possible contribution of learning based algorithms for MV imputation on such complex data sets should deserve further studies with a more powerful computational server. On the other hand, the QRILC method was invented for left-censored data imputation: MV are imputed by a random value generated by a truncated normal distribution. This method has been reported to better handle MNAR in metabolomics data sets than others [13]; the R wrapper based on "imputeLCMD" function has been retrieved from previous reported work [26].

2.5 FILTRATION STRATEGIES

The first filtration method assessed (named "t-test / fixed FC") relies on univariate statistic tests followed by the calculation of FC, for each ion. The Student test (t-test) is used since it is easy-to-use, it can handle rather small sample sets and moreover it has been already successfully implemented for untargeted food safety analysis [2,5]. Two successive t-tests are made, one between each group and the blank injections, the second between each group. The FC value for each ion is further calculated as the ratio between the median peak area (blanks and QC excluded) of the highest group over the median of the lowest. For each step (t-tests, FC), a fixed threshold is used for filtration (p-value < 0.05 for t-test, FC > 2 for fold change) whatever the ions considered.

The second approach (named "FC_{min}") is based on the calculation, for each ion, of the uncertainty on the FC [12] (U_{FC}), thanks to an error propagation estimation. U_{FC} enables then the determination of a relevant minimum fold change (FC_{min}) from which a significant effect can be distinguished from the overall method variability thanks to the equation: $FC_{min} = \frac{1}{1-U_{FC}}$. A peak is then selected if the corresponding FC is superior to the FC_{min}. Detailed calculation of FC_{min} can be found in the paper published by Ortmayr et al [12].

2.6 STUDY DESIGN AND METHODS PERFORMANCE

The different combinations of MV imputation/filtration methods (i.e. 6 different combinations, see Table 2) were tested on the previously described data sets, and the performance of each combination

assessed. Considering the study design and the fact that missing values were natively present in the tested data sets, limited quality metrics were available. Therefore, three indicators (2 quantitative, 1 qualitative) have been proposed to discuss the performance of the different combinations:

1. At the end of step B (Figure 1): total number of remaining ions after each combination, as well as number of ions of interest (i.e. "tracers") recovered. Venn diagrams have been used to spot the similarities and differences of selected ions between combinations;
2. At the end of step D (Figure 1): after multivariate analysis and annotation of suspect ions, group separation can be visually assessed and the detection percentage of the spiked contaminants ("tracers") can be calculated in both polarity modes. A global detection rate combining both ionization modes is determined as well.
3. Ease of implementation in the workflow (e.g. can the tool be implemented in-line with XCMS or the Matlab script or is a change of calculation platform needed? Is the method easy to handle?).

3. RESULTS AND DISCUSSION

3.1 STUDY OF MISSING VALUES IN DATA SETS

First of all, descriptive metrics on missing values were calculated on each data set for positive and negative mode, and their distribution in each data set visualized. These metrics include the global percentage of MV, group-wise missing value percentages, the respective rates of MAR and MNAR, Pearson correlation coefficients between MV rate for each ion and their *m/z*, RT or mean peak area. They are presented in Table 3. For the positive mode, a higher MV percentage is observed in data set #2 as compared to data set #1 (53.6 vs. 40.7%, respectively), but this phenomenon is not observed for the negative ionization mode.

To assess the presence of a trend within the distribution of missing values, a MV percentage is firstly calculated for each ion. Then, median of MV percentages is calculated for each percentile of relevant observed quantities (*m/z*, RT and median peak area of the ion), and the corresponding plots are drawn.

Illustrations for data set #1 in positive mode and data set #2 in negative mode are displayed in Supplementary material Figure S.1.

Strong similarities can be observed among our data sets. First of all, no correlation nor graphical trend could be established between the rate of MV and the measured *m/z* and RT, which might suggest that MV are distributed randomly regarding *m/z* and RT in our data sets (Pearson correlation coefficients between -0.14 and 0.13). Besides, even though the Pearson's correlation coefficients between MV rates and mean ion intensities are not significant (respectively 0.01 and -0.07), clear trends can be observed in the plot, with MV rates decreasing with the median peak area. Interestingly, when classifying missing values with our approach, MNAR were predominant (83.0-93.5% of total missing values) in all data sets, which is relevant with this trend, since, in MS data, MNAR often account for ions close to the limit of detection of the instrument [15]. A group-by-group study shows that they are distributed very evenly within the different sample groups, while the rate is lower in the QC samples (except for data set #2 in negative mode), and much higher in blank injections (this being expected). A slightly higher between-groups variability can be noticed for data set #2 in negative ionization mode, which cannot be explained.

Overall, the properties of both data sets, acquired using either negative or positive ionization modes, are very similar regarding missing values, even though less MV are observed in negative mode.

3.2 IONS SELECTION AFTER FILTRATION

For all combinations of missing value imputation and filtration methods two figures were monitored:

1. The total number of ions selected after filtration (meaning ions that pass pretreatment steps and will be used for multivariate data analysis);
2. The number of ions of interest selected.

Ions of interest for spiked contaminants (defined as $[M+H]^+$ and $[M+Na]^+$ forms for positive mode, $[M-H]^-$ for negative mode, and their corresponding M+1 and M+2 isotopic peaks) were *a posteriori* searched in the data matrices to assess any information loss during data treatments. Based on our spiking conditions (either 32 or 3 contaminants), a targeted screening of the initial data matrices reported a total of 57 ions of interest for our "tracers" in the data matrix built for data set #1, and 8 in the one built for

data set #2 for positive mode (respectively 36 and 4 in negative mode). Over the 57 ions of interest in positive mode, 54 have at least one missing value needing MV imputation to enable statistical selection (for negative mode: 36 over 36).

The effects of each filtration and MV imputation method on ions selection are visualized using Venn diagrams to spot common selected ions between method combinations (see Figure 3). For data set #1 in both polarity, with the t-test/fixed FC filtration, a common core of ions has been selected (871 for positive mode and 579 for negative mode) among which the majority of ions of interest (47/57 for positive mode, 36/36 for negative mode). This result highlights the ability of all MV imputation methods to allow the selection of relevant ions when combined with the t-test/fixed FC filtration method on these rather simple data sets (since 54/57 needed MV imputation for ESI^+ and 36/36 for ESI^-). The total number of ions selected in these data sets using FC_{min} filtration method is generally lower than with t-test/fixed FC (~100 vs. ~1,000, except for the negative mode with fillPeaks imputation). With FC_{min} filtration, fewer ions of interest are selected (14/57 for positive ionization mode and 29/36 for the negative mode); its influence on the final detection rate of the whole process will be discussed in 3.3.

The implementation of both filtration strategies on data set #2 (positive and negative ionization modes) leads to the selection of more ions than on data set #1. This is due to the higher between-samples variability, with about ten times more ions selected each time. As for data set #1, the common core of ions selected with the combinations containing the t-test/fixed FC filtration method (7,557 ions for ESI^+ and 8,964 for ESI^-) contains the majority of ions of interest (6/8 for positive mode and 4/4 for negative mode). The application of FC_{min} filtration on these data sets leads to the selection of less ions than for t-test/fixed FC as generally observed for data set #1. The common core of ions of interest is more reduced in positive mode (2/8) with combinations containing FC_{min} filtration. On this data set #2, combination containing SVD-QRILC method failed to recover the 4 ions of interest due to a too stringent filtration (about 1,500 ions selected against about 4,000 respectively). In negative mode, 3 ions of interest (out of 4) are selected with all combinations, and one extra-ion is picked by the fillPeaks method.

To conclude, considering the t-test/fixed FC filtration method, a common core of selected ions gathered the most part of ions of interest, meaning that all MV imputation methods are efficient to enable their selection during filtration. Comparison between MV imputation methods lies also in the number of total ions selected as this is indicative of the strength of the filtration. For FC_{min} filtration method, the conclusions are different since the common core of selected ions regroups less than 50% of ions of interest. In that case, fillPeaks and mean-LOD methods were more efficient for the selection of ions of interest, but in the meantime they led to high numbers of total ions selected.

3.3 GLOBAL PERFORMANCE OF THE APPROACH

The whole workflow (including the final multivariate and annotation steps) was considered to figure out which pretreatment method(s) offer(s) the best performances for untargeted food contaminants detection. Results for positive and negative ionization modes are presented in Table 4. A global detection rate of the method (obtained by the combination of results from both ionization modes) is displayed as well.

Firstly, for data set #1, whatever the pretreatment method combinations, all sample groups could be discriminated with our untargeted approach (see in Figure S.2 of supplementary materials for positive ionization mode score plots and Figure S.3 for negative ionization mode). It means that the multivariate method used can successfully separate the "unnecessary" ions in the data matrix from the common core of ions of interest observed both for t-test/fixed FC and FC_{min} (see in Table 4). Percentages of detection for our "tracers" ranged from 38 to 53% in positive mode and from 34 to 41% for negative mode (leading to global detection rates between 66 and 78% when combining both modes) for this data set that mimics a quite simple case (one brand, three different levels of contamination plus a control group). The influence of MV imputation method on the detection rates seems minor since all method combinations give acceptable performance. Interestingly, no clear link can be established between the number of ions of interest selected and the detection rate of the method, meaning that, even though FC_{min} filtration method selected less ions of interest than t-test/fixed FC (see Figure 3), it seems to select the most important ones (i.e. monoisotopic ions) with a minor impact on the detection rates observed.

Interestingly, for both positive and negative ionization modes, mean-LOD method coupled to FC_{min} filtration strategy seems to lead to lower relative intra-group variances (see in Figure S.2 and S.3 of supplementary materials). This may be the consequence of the use of the injection replicate information to fill missing values with this imputation methods, and also of the stronger data reduction brought by FC_{min} compared with t-test/fixed FC method.

On the other hand, all combinations do not seem suitable for the more complex data set #2. In positive ionization mode, mean-LOD & SVD-QRILC coupled with t-test/fixed FC filtration do not manage to detect the three contaminants spiked, and SVD-QRILC coupled with FC_{min} filtration only achieved the detection of 2 contaminants out of 3. Interestingly, all contaminants were detected using fillPeaks coupled with t-test/fixed FC or FC_{min} (combinations n°1 & 2 as displayed in Table 2) and mean-LOD coupled with FC_{min}. The performance of the methods are more homogeneous in negative ionization mode since only SVD/QRILC coupled with FC_{min} filtration failed to detect the spiked contaminants. At the end, when considering simultaneously both polarities, fillPeaks appears as the only MV imputation method that enables the annotation of all contaminants whatever the filtration method used.

Based on those results, as well as the easiness of implementation of each tool, the main characteristics of imputation and filtration methods were proposed (Table 5 and Table 6).

We observed that only combinations relying on fillPeaks successfully enabled the detection of spiked contaminants (or a majority of them) in all data sets. In addition, this MV imputation method does not need to classify missing value as MNAR or MAR, and it is easily implemented in-line after peak extraction since it is part of the XCMS package (being already implemented on every XCMS-based platforms). Practically speaking, fillPeaks is very user-friendly, with easy-to-use graphical interfaces developed by the community (e.g. W4M and XCMS Online). Yet, it relies on a complex algorithm, so that inconsistent results may be difficult to troubleshoot, especially for unexperienced users even though graphical outputs are available. Hopefully, this tool benefits from a very dynamic and open scientific community that brings help and technical support. As stated before, another drawback of this method, based on forced integration, lies in ions presenting flat baselines (cut-off during the acquisition) where

MVs are imputed as zeros and should be handled afterward since they may prevent the use of some critical pretreatment methods (e.g. log normalization and univariate statistics). This is illustrated by our results on data set MTBLS 771: a total of 48% of the values in the data matrix were missing before the fillPeaks step, while, after fillPeaks the data matrix contains 8% of zeros (i.e. 17% of initial missing values). This clearly shows that a significant number of zeros may be present after the fillPeaks completion, and suggests the advantage of combining fillPeaks with other MV imputation methods such as mean-LOD for example. The results obtained by the combination of both methods (fillPeaks and mean-LOD) can be found in the dedicated publication [27].

Methods needing MV classification suffer from the absence of established methodologies to classify missing values in MS-based data sets; under our experience, both methods (mean-LOD and SVD-QRILC) did not always enable the detection of contaminants in the most complicated case studied. Since MAR and MNAR are not imputed with the same algorithm, a MV misclassification may lead to an inconsistent imputation. The effect of such misclassification is expected to be higher with SVD-QRILC (based on statistical methods) than with the simple mean-LOD method, in agreement with the lower performance of SVD-QRILC compared with mean-LOD observed in this work. Our results pointed out the classification method for MAR and MNAR as a possible limiting step for the efficiency of these MVs imputation methods and further investigations discussing this first proposed methodology are needed. Despite this drawback, both methods have the advantage to fully complete the data matrix since no zeros are obtained at the end.

Regarding the mean-LOD method, the noise component set as a random value between -20% and +20% around the estimated values (mean or LOD) may sometimes over- or under-estimate the “real” standard deviation of the data. This over- or under-estimation may disturb the calculation of the FC uncertainty U_{FC} and influence the filtration method FC_{min}. However, the detection rate of the combination mean-LOD + FC_{min} indicated in Table 4 proved that the over- or under-estimation is not a critical issue for the tested data sets, but more tests are needed to confirm a larger applicability.

The t-test/fixed FC method relies on the use of fixed, generic threshold for each step (p-value < 0.05 for t-tests and FC > 2), which can be a limit since all ions do not necessarily have the same characteristics in terms of distribution and variance. On the opposite, the FC_{min} method adjusts the threshold to the measurement quality of each ion, which may enable a better quality of filtration, with the selection of peaks exhibiting lower relative standard deviations, and therefore potentially less artifacts. However, the t-test/fixed FC strategy offers more flexible applications than FC_{min} since one can use any univariate statistical test to better fit to the data structure, or apply one or two filtration steps (for example by omitting the fixed FC step) if too much data of interest seem to be lost. Even though FC_{min} leads to the selection of fewer ions of interest than t-test/fixed FC, the global detection rates obtained are very similar (see Table 4). On more complex data set such as data set #2, the greater reduction of ions number generally observed with the FC_{min} filtration method can also be an asset since it makes the computation easier and faster. In the meantime, the risk of discarding a potential contaminant is also higher, especially with molecules having a signal close to the limit of detection of the instrument. Consequently, it could be recommended to implement both filtration methods in parallel to increase the detection probability of potential contaminants.

4. CONCLUSION

Several pretreatment methods (three missing value imputation methods - one based on the forced integration of raw data, two based on the classification of missing values as MAR or MNAR - coupled to two filtration methods, leading to six combinations) were tested on two LC-MS data sets dedicated to untargeted food chemical safety. They were integrated in a general workflow, and the final detection rate calculated for each data set and method combination. In addition to this global performance assessment, the ions selected by each combination were more deeply investigated.

As expected initially, the total number of ions selected varies a lot between pretreatment methods. Interestingly the ions of interest (corresponding to spiked contaminants) were selected by most methods. Considering the whole workflow, all combinations were able to detect the spiked contaminants on the data sets corresponding to a simple contamination scenario (positive and negative ionization modes),

with different success rates (from 66 to 78%). The more heterogeneous data set was more problematic since several combinations did not enable the detection of the spiked contaminants. In fact, the only imputation method that enables the detection of our tracers for this contamination scenario, whatever the filtration approach used, is fillPeaks based on the re-analysis of raw data. This tools has also the advantage to be easily implemented in-line with the peak extraction step if this one is carried out with the wide-spread, user friendly package XCMS or its online implementations XCMS-online or Workflow4Metabolomics. However, on data exhibiting a flat baseline with no signal in case of no peak, it can generate an important amount of zeros. In that case, they should be handled as missing values to avoid any problematic issues in the workflow afterwards. We suggest that mean-LOD method should be used to complement fillPeaks on remaining zeros since it is very easy to implement and still shows satisfactory results.

Unlike existing missing value imputation approaches, two methods presented here rely on a classification of missing values according to their nature. This very simple methodology is based on instrumental replicates, thereby authorizing a quick classification; in addition, it can be easily combined with any MV imputation method. Yet, it seems to face some limits when dealing with heterogeneous data sets, so that more work is needed to better address MV classification for MS-based data sets. In this work, this classification-based approach has been used with either a simple method (mean-LOD) or a more sophisticated one chosen for its performances on respective missing values types (SVD-QRILC). The results presented here constitute a good proof of concept of the potential of such classification-based approaches to help missing value imputation. It would surely benefits from its implementation with other imputation methods such as ones based on machine learning algorithms, for example Artificial Neural Networks (ANN, [28]) or genetic algorithm [29]. Such work would be a natural extension of the present publication and would provide highly interesting results for the scientific community, even outside the field of untargeted food safety assessment. In addition, studying the proposed missing value imputation strategy on a simulated data set (i.e. on better controlled, even though less realistic, situation) could lead to interesting contribution to the understanding of the missing value imputation process.

In applied fields such as untargeted food chemical safety assessment, the user mainly focus on the final outcome of the approach, but our understanding of the process should be improved in order to build better tools and workflows. This work shows a first attempt in that direction but more work and more data sets dealing with untargeted food safety are needed to get a critical point of view on all the steps of the workflow and their influence on the detection rates.

FUNDING

This work was supported by Paris Institute of Technology for Life, Food and Environmental Sciences (AgroParisTech), the French National Institute for Agricultural Research (INRA) and the French Ministry of Higher Education and Research.

BIBLIOGRAPHY

[1] E. Tengstrand, J. Rosén, K.E. Hellenäs, K.M. Åberg, A concept study on non-targeted screening for chemical contaminants in food using liquid chromatography-mass spectrometry in combination with a metabolomics approach, Anal. Bioanal. Chem. 405 (2013) 1237–1243. doi:10.1007/s00216-012-6506-5.

[2] A.M. Knolhoff, J.A. Zweigenbaum, T.R. Croley, Nontargeted Screening of Food Matrices: Development of a Chemometric Software Strategy to Identify Unknowns in Liquid Chromatography-Mass Spectrometry Data, Anal. Chem. 88 (2016) acs.analchem.5b04208. doi:10.1021/acs.analchem.5b04208.

[3] J. Cotton, F. Leroux, S. Broudin, M. Marie, B. Corman, J.C. Tabet, C. Ducruix, C. Junot, High-resolution mass spectrometry associated with data mining tools for the detection of pollutants and chemical characterization of honey samples, J. Agric. Food Chem. 62 (2014) 11335–11345. doi:10.1021/jf504400c.

[4] M. Kunzelmann, M. Winter, M. Åberg, K.-E. Hellenäs, J. Rosén, Non-targeted analysis of

unexpected food contaminants using LC-HRMS, Anal. Bioanal. Chem. (2018) 1–10. doi:10.1007/s00216-018-1028-4.

[5] G. Delaporte, M. Cladière, D. Jouan-Rimbaud Bouveresse, V. Camel, Untargeted food contaminant detection using UHPLC-HRMS combined with multivariate analysis: Feasibility study on tea, Food Chem. 277 (2019) 54–62. doi:10.1016/j.foodchem.2018.10.089.

[6] W.B. Dunn, W. Lin, D. Broadhurst, P. Begley, M. Brown, E. Zelena, A.A. Vaughan, A. Halsall, N. Harding, J.D. Knowles, S. Francis-McIntyre, A. Tseng, D.I. Ellis, S. O'Hagan, G. Aarons, B. Benjamin, S. Chew-Graham, C. Moseley, P. Potter, C.L. Winder, C. Potts, P. Thornton, C. McWhirter, M. Zubair, M. Pan, A. Burns, J.K. Cruickshank, G.C. Jayson, N. Purandare, F.C.W. Wu, J.D. Finn, J.N. Haselden, A.W. Nicholls, I.D. Wilson, R. Goodacre, D.B. Kell, Molecular phenotyping of a UK population: defining the human serum metabolome, Metabolomics. 11 (2014) 9–26. doi:10.1007/s11306-014-0707-1.

[7] E.A. Thévenot, A. Roux, Y. Xu, E. Ezan, C. Junot, Analysis of the Human Adult Urinary Metabolome Variations with Age, Body Mass Index, and Gender by Implementing a Comprehensive Workflow for Univariate and OPLS Statistical Analyses, J. Proteome Res. 14 (2015) 3322–3335. doi:10.1021/acs.jproteome.5b00354.

[8] J.P. Antignac, F. Courant, G. Pinel, E. Bichon, F. Monteau, C. Elliott, B. Le Bizec, Mass spectrometry-based metabolomics applied to the chemical safety of food, TrAC - Trends Anal. Chem. 30 (2011) 292–301. doi:10.1016/j.trac.2010.11.003.

[9] M. Castro-Puyana, R. Pérez-Míguez, L. Montero, M. Herrero, Application of mass spectrometry-based metabolomics approaches for food safety, quality and traceability, TrAC - Trends Anal. Chem. 93 (2017) 102–118. doi:10.1016/j.trac.2017.05.004.

[10] A.M. Knolhoff, T.R. Croley, Non-targeted screening approaches for contaminants and adulterants in food using liquid chromatography hyphenated to high resolution mass spectrometry, J. Chromatogr. A. 1428 (2016) 86–96. doi:10.1016/j.chroma.2015.08.059.

[11] C. Roullier, Y. Guitton, M. Valery, S. Amand, S. Prado, T. Robiou Du Pont, O. Grovel, Y.F. Pouchus, Automated Detection of Natural Halogenated Compounds from LC-MS Profiles-Application to the Isolation of Bioactive Chlorinated Compounds from Marine-Derived Fungi, Anal. Chem. 88 (2016) 9143–9150. doi:10.1021/acs.analchem.6b02128.

[12] K. Ortmayr, V. Charwat, C. Kasper, S. Hann, G. Koellensperger, Uncertainty budgeting in fold change determination and implications for non-targeted metabolomics studies in model systems, Analyst. 142 (2017) 80–90. doi:10.1039/C6AN01342B.

[13] R. Wei, J. Wang, M. Su, E. Jia, S. Chen, T. Chen, Y. Ni, Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data, Sci. Rep. 8 (2018) 663. doi:10.1038/s41598-017-19120-0.

[14] O. Hrydziuszko, M.R. Viant, Missing values in mass spectrometry based metabolomics: An undervalued step in the data processing pipeline, Metabolomics. 8 (2012) 161–174. doi:10.1007/s11306-011-0366-4.

[15] C. Lazar, L. Gatto, M. Ferro, C. Bruley, T. Burger, Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies, J. Proteome Res. 15 (2016) 1116–1125. doi:10.1021/acs.jproteome.5b00981.

[16] R. Di Guida, J. Engel, J.W. Allwood, R.J.M. Weber, M.R. Jones, U. Sommer, M.R. Viant, W.B. Dunn, Non-targeted UHPLC-MS metabolomic data processing methods: a comparative investigation of normalisation, missing value imputation, transformation and scaling, Metabolomics. 12 (2016). doi:10.1007/s11306-016-1030-9.

[17] C.A. Smith, E.J. Want, G. O'Maille, R. Abagyan, G. Siuzdak, XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification, Anal. Chem. 78 (2006) 779–787. doi:10.1021/ac051437y.

[18] M. Cladière, G. Delaporte, E. Le Roux, V. Camel, Multi-class analysis for simultaneous determination of pesticides, mycotoxins, process-induced toxicants and packaging contaminants

in tea, Food Chem. 242 (2018) 113–121. doi:10.1016/j.foodchem.2017.08.108.

[19] K. Haug, R.M. Salek, P. Conesa, J. Hastings, P. De Matos, M. Rijnbeek, T. Mahendraker, M. Williams, S. Neumann, P. Rocca-Serra, E. Maguire, A. González-Beltrán, S.A. Sansone, J.L. Griffin, C. Steinbeck, MetaboLights - An open-access general-purpose repository for metabolomics studies and associated meta-data, Nucleic Acids Res. 41 (2013) 781–786. doi:10.1093/nar/gks1004.

[20] M.C. Chambers, B. MacLean, R. Burke, D. Amodei, D.L. Ruderman, S. Neumann, L. Gatto, B. Fischer, B. Pratt, J. Egertson, K. Hoff, D. Kessner, N. Tasman, N. Shulman, B. Frewen, T.A. Baker, M.Y. Brusniak, C. Paulse, D. Creasy, L. Flashner, K. Kani, C. Moulding, S.L. Seymour, L.M. Nuwaysir, B. Lefebvre, F. Kuhlmann, J. Roark, P. Rainer, S. Detlev, T. Hemenway, A. Huhmer, J. Langridge, B. Connolly, T. Chadick, K. Holly, J. Eckels, E.W. Deutsch, R.L. Moritz, J.E. Katz, D.B. Agus, M. MacCoss, D.L. Tabb, P. Mallick, A cross-platform toolkit for mass spectrometry and proteomics, Nat. Biotechnol. 30 (2012) 918–920. doi:10.1038/nbt.2377.

[21] F. Giacomoni, G. Le Corguillé, M. Monsoor, M. Landi, P. Pericard, M. Pétéra, C. Duperier, M. Tremblay-Franco, J.F. Martin, D. Jacob, S. Goulitquer, E.A. Thévenot, C. Caron, Workflow4Metabolomics: A collaborative research infrastructure for computational metabolomics, Bioinformatics. 31 (2015) 1493–1495. doi:10.1093/bioinformatics/btu813.

[22] R. Tautenhahn, C. Bottcher, S. Neumann, Highly sensitive feature detection for high resolution LC/MS, BMC Bioinformatics. 9 (2008) 16. doi:10.1186/1471-2105-9-504.

[23] D.N. Rutledge, D. Jouan-Rimbaud Bouveresse, Corrigendum to “Independent Components Analysis with the JADE algorithm”, [Analytical Chemistry, 50, (2013) 22-32, doi:10.1016/j.trac.2013.03.013], TrAC - Trends Anal. Chem. 67 (2015) 220. doi:10.1016/j.trac.2015.02.001.

[24] G. Libiseller, M. Dvorzak, U. Kleb, E. Gander, T. Eisenberg, F. Madeo, S. Neumann, G. Trausinger, F. Sinner, T. Pieber, C. Magnes, IPO: a tool for automated optimization of XCMS

parameters, BMC Bioinformatics. 16 (2015) 118. doi:10.1186/s12859-015-0562-8.

[25] W. Stacklies, H. Redestig, M. Scholz, D. Walther, J. Selbig, pcaMethods - A bioconductor package providing PCA methods for incomplete data, Bioinformatics. 23 (2007) 1164–1167. doi:10.1093/bioinformatics/btm069.

[26] C. Lazar, imputeLCMD: A collection of methods for left-censored missing data imputation v.2.0, 2015. https://cran.r-project.org/package=imputeLCMD.

[27] G. Delaporte, M. Cladière, V. Camel, Untargeted food chemical safety assessment : A proof-of-concept on two analytical platforms and contamination scenarios of tea, Food Control. 98 (2019) 510–519. doi:10.1016/j.foodcont.2018.12.004.

[28] E.G. Armitage, J. Godzien, V. Alonso-Herranz, Á. López-Gonzálvez, C. Barbas, Missing value imputation strategies for metabolomics data, Electrophoresis. 36 (2015) 3050–3060. doi:10.1002/elps.201500352.

[29] I.B. Aydilek, A. Arslan, A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm, Inf. Sci. (Ny). 233 (2013) 25–35. doi:10.1016/j.ins.2013.01.021.

552 TABLE 1 MAIN CHARACTERISTICS OF STUDIED DATA SETS [5]

Data set	Number of brands	Spiking mix*	Spiking levels (µg/kg)
#1	1	32 contaminants	0; 10; 50; 100
#2	2	3 contaminants	0; 50 (for each brand)

553 ** Details on spiked contaminants can be found in Supplementary material - Table S.1*

554

555 TABLE 2 COMBINATION OF PRETREATMENT METHODS TESTED

Combination n°	Missing value imputation method	Filtration method
1	fillPeaks	t-test/fixed FC
2	fillPeaks	FC_{min}
3	Mean-LOD	t-test/fixed FC
4	Mean-LOD	FC_{min}
5	SVD-QRILC	t-test/fixed FC
6	SVD-QRILC	FC_{min}

556

557

558 TABLE 3 PROPERTY SUMMARY OF DATA SETS (FOR BOTH IONIZATION MODES)

		Data set #1		Data set #2	
		POS	**NEG**	**POS**	**NEG**
Number of extracted ions (XCMS)		29,755	24,543	23,891	17,269
Number of data files		48	48	57*	59
Global MV rate		40.7	30.1	53.6	23.3
Group-wise MV rates	Blanks	92.7	95.3	94.6	93.8
	QC	29.7	17.0	43.7	21.5
	Group 1	32.4	19.9	50.1	11.6
	Group 2	31.9	19.3	49.2	12.5
	Group 3	32.8	19.3	51.2	18.9
	Group 4	31.2	18.2	51.3	21.6
MNAR % in missing values		86.3	83.0	93.5	84.0
MAR % in missing values		13.7	17.0	6.5	16.0
Pearson correlation MV / *m/z*		0.02	-0.01	0.13	0.02
Pearson correlation MV / RT		0.08	-0.14	0.07	-0.07
Pearson correlation MV / mean area		0.01	-0.06	-0.18	-0.07

559 ** Injections outliers were visually detected in data set #2 for positive mode, and thus discarded*

560

561 TABLE 4 PERFORMANCES OF THE WHOLE WORKFLOW DEPENDING ON THE MISSING VALUE

562 IMPUTATION / FILTRATION METHODS COMBINATION AND DATA SETS

Data set	**Combination**	**Positive mode**		**Negative mode**		**Global**
		Number of ions after filtration	Detection rate (%)	Number of ions after filtration	"Tracers" detection (%)	Detection rate (%)
#1	fillPeaks + t-test/fixed FC	1,710	44	952	38	**66**
	fillPeaks + FC_{min}	328	50	2,136	38	**72**
	mean-LOD + t-test/fixed FC	4,309	50	2,950	41	**75**
	mean-LOD + FC_{min}	210	44	552	34	**66**
	SVD-QRILC + t-test/fixed FC	3,336	53	2,607	38	**78**
	SVD-QRILC + FC_{min}	160	38	403	38	**66**
#2	fillPeaks + t-test/fixed FC	9,778	100	9,381	67	**100**
	fillPeaks + FC_{min}	4,572	100	9,383	67	**100**
	mean-LOD + t-test/fixed FC	14,142	0	11,261	67	**67**
	mean-LOD + FC_{min}	3,530	100	5,238	67	**100**
	SVD-QRILC + t-test/fixed FC	13,188	0	11,066	67	**67**
	SVD-QRILC + FC_{min}	1,524	67	4,060	0	**67**

563

564 TABLE 5 MAIN FEATURES OF MISSING VALUE IMPUTATION METHODS

MV imputation method	Main characteristics
fillPeaks	*Pros* No need for MV classification Easy in-line implementation within XCMS Gives good results with all filtration methods on every data sets Benefits from the support of a dynamic scientific community Can be easily combined with other statistical MV imputation methods *Cons* May generate a lot of zeros on flat baseline with no signal (i.e. issues with log scaling and univariate statistics) Relies on a complex algorithm that can be difficult to troubleshoot, especially for unexperienced users even though graphical outputs are available
Mean-LOD	*Pros* Simple tools, understandable by all No zeros at the end of the process *Cons* Needs MV classification (= more complex to implement and may be subjected to discussion) May lead to over-fitting of the data Does not enable the detection of all "tracers" when combined with t-test / fixed FC on the most heterogeneous data set
SVD-QRILC	*Pros* Best detection rate on the simple data set No zeros at the end of the process *Cons* Needs MV classification More complex methods than mean-LOD, may be difficult to troubleshoot for unexperienced users Performs badly on the most heterogeneous data set

565

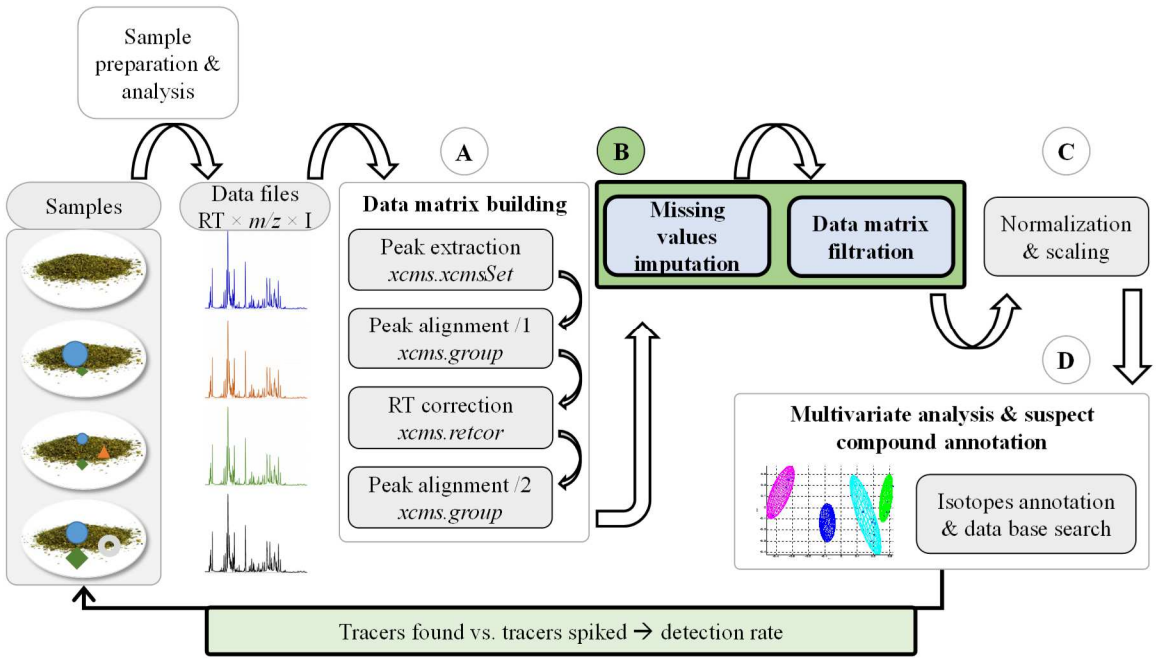
566

567 TABLE 6 MAIN FEATURES OF FILTRATION METHODS

Filtration method	**Main characteristics**
t-test/fixed FC	*Pros* Easy to implement More ions of interest selected *Cons* Lower detection rates on heterogeneous data sets with imputation methods other than fillPeaks Filtration thresholds may be subjected to discussion
FC_{min}	*Pros* Easy to implement No parameter to set (filtration threshold determined by the quality of measurement for each ion) Greater reduction of selected ion numbers Goes well with most imputation methods used *Cons* Fewer ions of interest selected

568

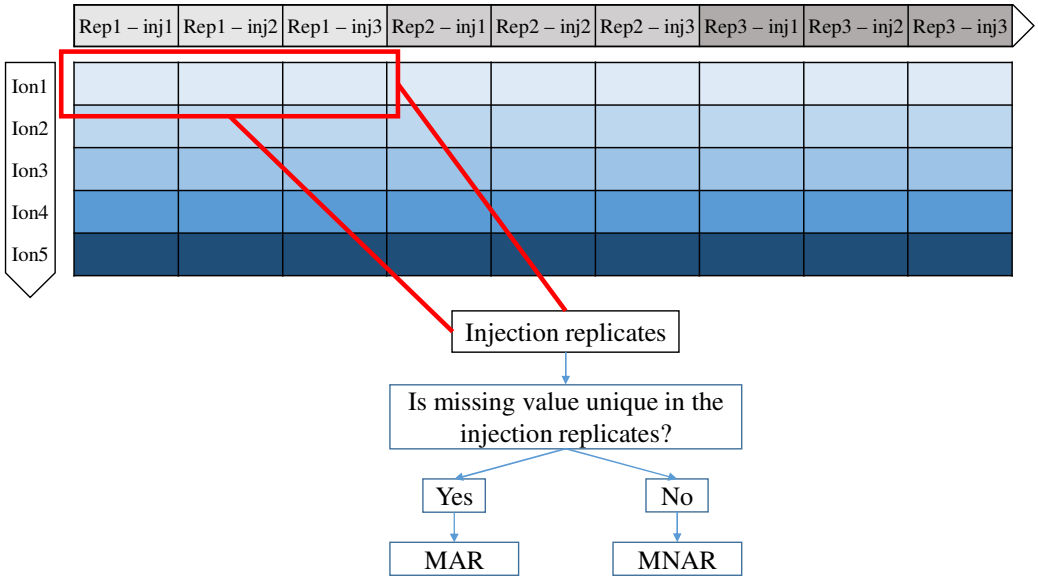
569



570

571 FIGURE 1 WORKFLOW IMPLEMENTED

572



573

574 FIGURE 2 DATA MATRIX LAYOUT AND REPRESENTATION OF THE CLASSIFICATION

575 METHODOLOGY FOR MISSING VALUES

576

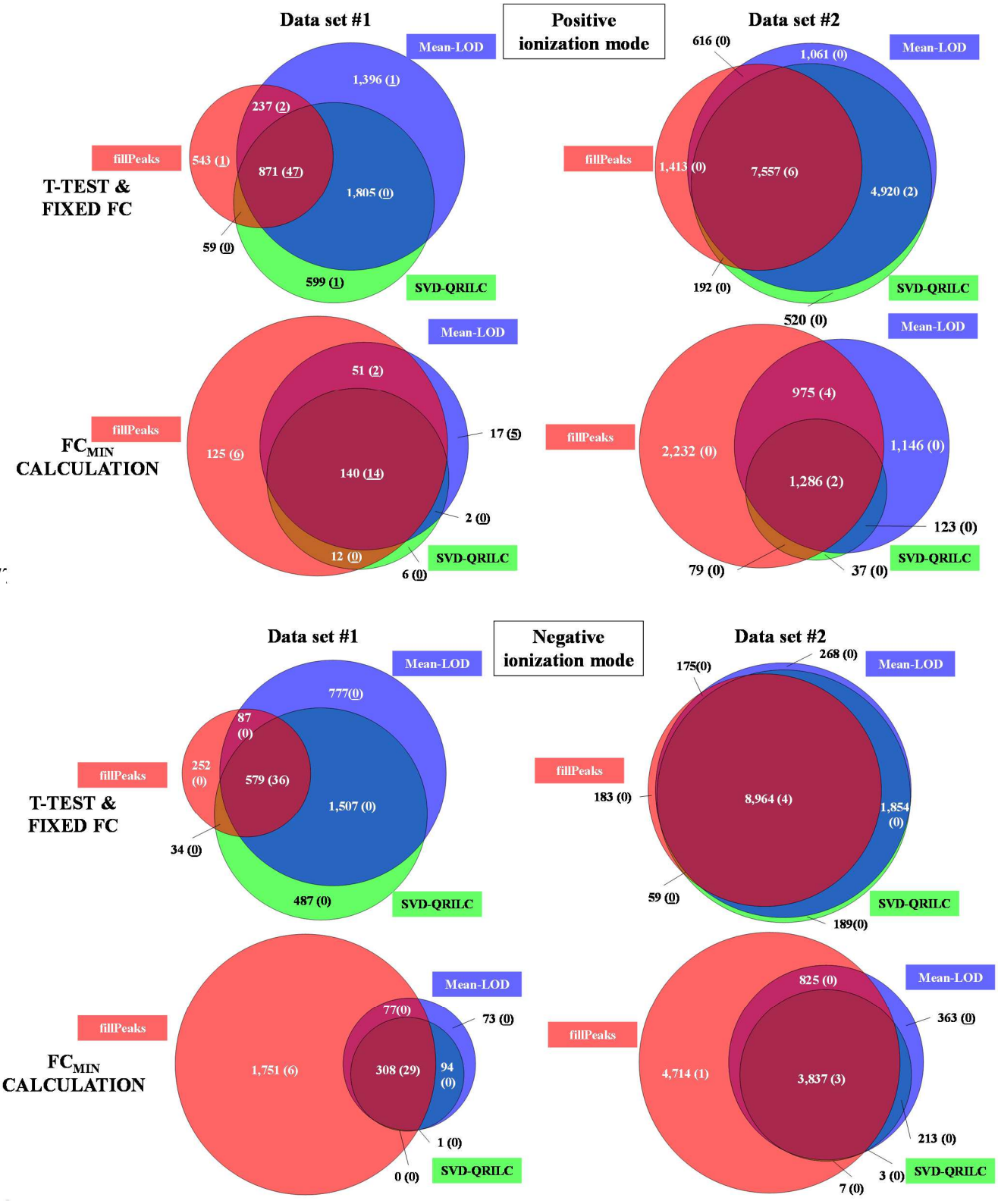


FIGURE 3 VENN DIAGRAMS FOR NUMBER OF IONS* SELECTED BY EACH PRETREATMENT COMBINATION ON BOTH DATA SETS FOR POSITIVE AND NEGATIVE IONIZATION MODES

*number of ions relative to our "tracers" are indicated within parenthesis