Received: 13 October 2017 | Revised: 1 December 2017 | Accepted: 7 December 2017

DOI: 10.1002/cem.3001

SPECIAL ISSUE - RESEARCH ARTICLE

WILEY Journal of CHEMOMETRICS

Significance of variables for discrimination: Applied to the search of organic ions in mass spectra measured on cometary particles

Kurt Varmuza[1] | Peter Filzmoser[1] | Irene Hoffmann[1] | Jan Walach[1] | Hervé Cottin[2] | Nicolas Fray[2] | Christelle Briois[3] | Paola Modica[3] | Anaïs Bardyn[4] | Johan Silén[5] | Sandra Siljeström[6] | Oliver Stenzel[7] | Jochen Kissel[7] | Martin Hilchenbach[7]

[1] Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology, Vienna, Austria

[2] Université Paris Est Créteil et Université Paris Diderot, LISA, UMR CNRS 7583, Institut Pierre Simon Laplace, Créteil, France

[3] Laboratoire de Physique et Chimie de l'Environnement et de l'Espace (LPC2E), CNRS, Université d'Orléans, Orléans, France

[4] Department Terrestrial Magnetism, Carnegie Institution of Washington, Washington, DC, USA

[5] Finnish Meteorological Institute, Helsinki, Finland

[6] RISE Research Institutes of Sweden, Bioscience and Materials/Chemistry and Materials, Stockholm, Sweden

[7] Max-Planck-Institute for Solar System Research, Göttingen, Germany

Correspondence
Kurt Varmuza, Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology, Vienna, Austria.
Email: kurt.varmuza@tuwien.ac.at

Funding information
Austrian Science Fund, Grant/Award Number: P 26871-N20

Abstract

The instrument Cometary Secondary Ion Mass Analyzer (COSIMA) on board of the European Space Agency mission Rosetta to the comet 67P/Churyumov-Gerasimenko is a secondary ion mass spectrometer with a time-of-flight mass analyzer. It collected near the comet several thousand particles, imaged them, and analyzed the elemental and chemical compositions of their surfaces. In this study, variables have been generated from the spectral data covering the mass ranges of potential C-, H-, N-, and O-containing ions. The variable importance in binary discriminations between spectra measured on cometary particles and those measured on the target background has been estimated by the univariate *t* test and the multivariate methods discriminant partial least squares, random forest, and a robust method based on the log ratios of all variable pairs. The results confirm the presence of organic substances in cometary matter—probably a complex macromolecular mixture.

KEYWORDS

classification, comet 67P/Churyumov-Gerasimenko, D-PLS, pairwise log ratios, random forest, Rosetta, time-of-flight secondary ion mass spectrometry, variable importance

1 | INTRODUCTION

The Rosetta spacecraft of the European Space Agency was launched on 2 March 2004 and reached the comet 67P/Churyumov-Gerasimenko (short 67P or nicknamed *Chury*[1]) at a distance of about 100 km on 6 August 2014. Rosetta escorted the comet at distances between a few kilometers and 1500 km on its orbit around the Sun for more than 2 years.

 Copyright © 2018 John Wiley & Sons, Ltd.

The mission was terminated by a controlled touchdown on the comet on 30 September 2016. A number of instruments on Rosetta investigated the comet and the coma around it. On board the Rosetta orbiter, the instrument Cometary Secondary Ion Mass Analyzer (COSIMA) collected cometary dust particles, imaged them, and analyzed their surfaces by time-of-flight secondary ion mass spectrometry (TOF-SIMS). This work evaluates a set of positive ion mass spectra with the aim to obtain information about ions consisting of C, H, N, and O atoms, originating from presumable organic material in the refractory part of the cometary particles. The data analysis methods applied are mostly from multivariate data analysis (chemometrics) and have been used to estimate the importance of variables (which are related to ion masses and ion formulae) for a discrimination of spectra measured on cometary particles and on the background.

We first give a brief overview about the comet 67P, the COSIMA instrument, the collected particles, and the specific properties of the available measurement data. Then the selection of data (mass spectral intervals with secondary ion counts) for CHNO ions is described, and the transformation into data matrices for spectra measured on cometary particles and for spectra measured on the gold target background used for the collection. The applied complementary methods for discriminating between these 2 classes are outlined, namely, *t* test, discriminant partial least squares (D-PLS), and random forest (RF), as well as the recently suggested robust method based on pairwise log ratios of the variables (rPLR). The results are various criteria for the variable importance for class discrimination that may indicate CHNO ions originating from the investigated cometary material.

2 | EXPERIMENTAL

2.1 | Comet 67P and Rosetta mission

Comet 67P moves around the Sun in an orbit with a maximum distance (aphelion) to the Sun of 5.7 AU and a minimum distance (perihelion) to the Sun of 1.24 AU, with a period of 6.44 years. The unit AU stands for *astronomical unit*, defined as a distance of 149 597 870 700 m, which is approximately the mean distance between the Earth and Sun. The size of 67P is approximately 6 km × 4 km × 3 km with an irregular shape; the rotation period is 12.4 hours, mean density 0.533 g/cm^3, and light reflectance (albedo) 6% (thus appearing very black). The surface of the comet is highly diverse, including sandy flat areas, boulders (some 10 m), canyons, and steep cliffs (more than 100 m). Several pits (sinkholes) exhibited outbursts of gases and dust, especially during the perihelion phase.

The Rosetta spacecraft[2-4] had a size of about 2.8 m × 2.1 m × 2 m with a mass of about 3 tons. It carried about 1.7-ton propellant (methylhydrazine and N_2O_4); electrical power was produced by a 32-m-wide solar panel; communication with ground was provided by a 2.2-m dish antenna. Eleven instruments on Rosetta investigated the nucleus, gas, and particles (especially by the COSIMA instrument, Section 2.2), and solar wind interactions. The lander *Philae* reached the comet surface on 12 November 2014 and recorded mass spectra that indicate a mixture of low–molecular weight substances (up to molar mass 62), present in the gas phase near the surface of the comet. A comparative study of the mass spectra from 3 instruments (on the lander and the orbiter) revealed the presence of a variety of CHO molecules, together with minor amounts of N- and S-containing substances up to 92 molar mass.[5]

2.2 | Instrument COSIMA

The instrument COSIMA[6] on board Rosetta consisted of 3 main parts: (1) a target storage and manipulation unit; (2) a microscope with camera (COSISCOPE); and (3) a TOF-SIMS. The mass of COSIMA was 20 kg, and its power consumption was 20 W.

The cometary particles were collected on a set of gold or silver targets of size 1 cm × 1 cm. Three targets were mounted together on each target holder; a total of 24 target holders were available. The target holder was transported to the different instrument subsystems—such as the dust collection funnel, the microscope, and the mass spectrometer—by a robotic device.[6]

The COSISCOPE microscope[7] delivered images from the targets with 1024 × 1024 pixels (each 14 μm × 14 μm) with a grazing incidence illumination of 2 LEDs from opposite sides. Evaluation of the images was performed on ground, basically the identification of collected particles and the determination of their *x* and *y* coordinates and size.

The TOF-SIMS used greater than 99.9% isotope-pure 115-indium for the production of primary ions (energy 8 keV, 3-ns pulses with about 1000 ions, repetition rate 1.5 kHz). The area (footprint) hit by the primary ion beam was about 35 μm × 50 μm (full width at half maximum intensity). The secondary ions were extracted by 2 kV into a 54-cm drift tube at 1 kV; a 2-stage ion reflector compensated for varying kinetic energies of ions with identical mass-to-charge ratios

(m/z). After passing a 51-cm drift tube, the ions were counted by a detector, and the flight time was measured with a resolution of 1 nanosecond. A typical mass spectrum required approximately 2.5-minute acquisition time with 225 000 primary ion shots.

The m/z is related to the flight time t by $t = a + b(m/z)^{0.5}$; instrumental parameters a and b define the mass calibration. The mass resolution was about $m/\Delta m = 1000$ at m/z 73 (Δm is the full width at half maximum peak height) and 500 at m/z 12. This mass resolution allows a separation of elemental ions from H-rich ions of the same nominal (integer) mass in this mass range. However, a complete separation of the different, principally possible CHNO ions of the same nominal mass is in general not possible (Section 2.5).

The measurement data of a mass spectrum comprise the ion counts for 128 000 time bins (width 1.95 ns) for the full mass range of up to 6500 Da and 42 000 time bins for the mass range 0 to 300 Da. Full data of more than 30 000 mass spectra have been sent to ground.

2.3 | Cometary particle samples

The typical exposure time for particle collection was between 1 day and 1 week; distance to the comet surface was typically 10 to 200 km and distance to the Sun 1.2 to 4.7 AU. In total, during the more than 2 years next to the comet, COSIMA collected about 1400 particles or more than 30 000 particle fragments within 538 days of net exposure time.[7-9] Sizes of the evaluated particles are between about 14 μm (pixel size) and 1000 μm; density was estimated at 0.2 to 0.3 g/cm^3, and the porosity as greater than 80%.[10] Most particles have an agglomerative, fluffy, porous structure. The impact speed of the particles was a few meters per second and thus did not affect the chemical composition—contrary to the sampling with several kilometers per second during the Stardust mission with samples returned to Earth.[11,12] On about 250 selected particles, TOF-SIMS spectra have been measured, showing a mixture of elements close to the chondritic meteorite composition but enriched in Si and C.[13,14] Particles are made in mass of about 55% silicates and 45% carbonaceous material.[13] The atomic ratio C/Si was estimated at 5.5 (±1.3),[13] the atomic ratio N/C was estimated at 0.035 (±0.011),[15] and the carbon may be mostly present in macromolecular substances.[16]

2.4 | Mass spectral data

The spectra used in this work have been measured on 2 particles and on the gold background of the targets used for particle collection; for a summary, see Table 1. One particle—named *Sai*[9]—had a diameter of about 80 μm and was collected on 13/14 April 2016 on a target covered with porous Au (black, named *3D1*).[17] The *Sai* data set contains 29 spectra. The other particle—named *Kerttu*—had a diameter of about 400 μm and was collected on 25/26 March 2016 on a target also covered with porous Au (black, named *3D0*); 23 spectra are in data set *Kerttu*. A third set with 59 spectra has been measured on a particle-free region of target *3D0* in the neighborhood of *Kerttu*, forming the data set background.

The primary selection of particles and mass spectra considered a sufficient size of the particle, high signals for $^{56}Fe^+$, $^{24}Mg^+$, and $CH_{0\text{-}3}^+$.[13] A further selection of the spectra actually measured on a particle was necessary because of an uncertainty of about ±50 μm of the position of the primary ion beam. For this purpose, 1-class classification methods[18] have been applied, with the single class defined by the background spectra. A successful approach was modeling the background spectra by a robust principal component analysis (PCA) method.[19] A spectrum has been considered as being measured on a particle if the orthogonal distance and the score distance from the PCA model are both large.[20]

TABLE 1 Mass spectral data sets

Type	Particle	Target	No. spectra
Comet	*Sai*	*3D1*	29
Comet	*Kerttu*	*3D0*	23
Background	...	*3D0*	59

2.5 | Preprocessing of mass spectra

2.5.1 | Basic spectral data treatment

The mass range considered in this work is from 12 Da ($^{12}C^+$ with mass 12.000) to 72 Da (the CHNO ion with maximum mass 72.094 is $^{12}C_5H_{12}^+$). The upper limit was guided (1) by the presence of a high peak at mass 73.047 from $Si(CH_3)_3^+$ due to the instrument/spacecraft contamination by polydimethylsiloxane (PDMS); (2) by the low intensities of CHNO ions with mass >15, accompanied with an increasing number of formula assignments at higher masses; and (3) by the insufficient separation of CHNO ions from inorganic/elemental ions at higher masses. The raw spectral data have been mass calibrated in the range 0 to 700 Da, with the ion count data mapped to a fixed raster of constant flight time intervals (rebinning[14,21]). For instance, the intervals (mass bins) around mass 12 are defined by the masses 11.99396, 11.99833, 12.00270, and 12.00707 (width 0.00437) and around mass 73 by the masses 72.98766, 72.99844, 73.00922, and 73.02000 (width 0.01078).

Furthermore, the mass scale of each spectrum has been individually adjusted with the reference ions $^{12}C^+$, $^{23}Na^+$, and $^{28}Si(^{12}CH_3)_3^+$. Gauss peaks have been fitted to the experimental ion count data for these ions. The standard deviation s of a Gauss peak is related to the mass resolution by $s = \Delta m/2.355$, with Δm for the full width at half maximum peak height. The mean (peak mass) and the maximum (peak height) of a Gauss peak are optimized by using the function *optim*() in the programming environment R.[22] The resulting errors of the reference masses for the used data are between ±0.001 Da for mass 12 and ±0.005 Da for mass 73. Then the mass scales of the spectra have been adjusted for 0 error at the reference masses and by a linear interpolation between them. Finally, the ion count data have been again mapped to the original raster of mass bins.

Examples of recalibrated spectra are shown in Figure 1. The mass calibration of the background spectra is less stable than that of the particle spectra. The main reasons for the appearing errors of about ±1 bin are probably the low ion counts for the reference Na^+ in the background spectra. Maximum ion counts per bin of these peaks are 4 to 13, causing some irregular shapes of the peaks. Tests with simulated errors in the spectra showed that mass scale errors of ±1 bin do

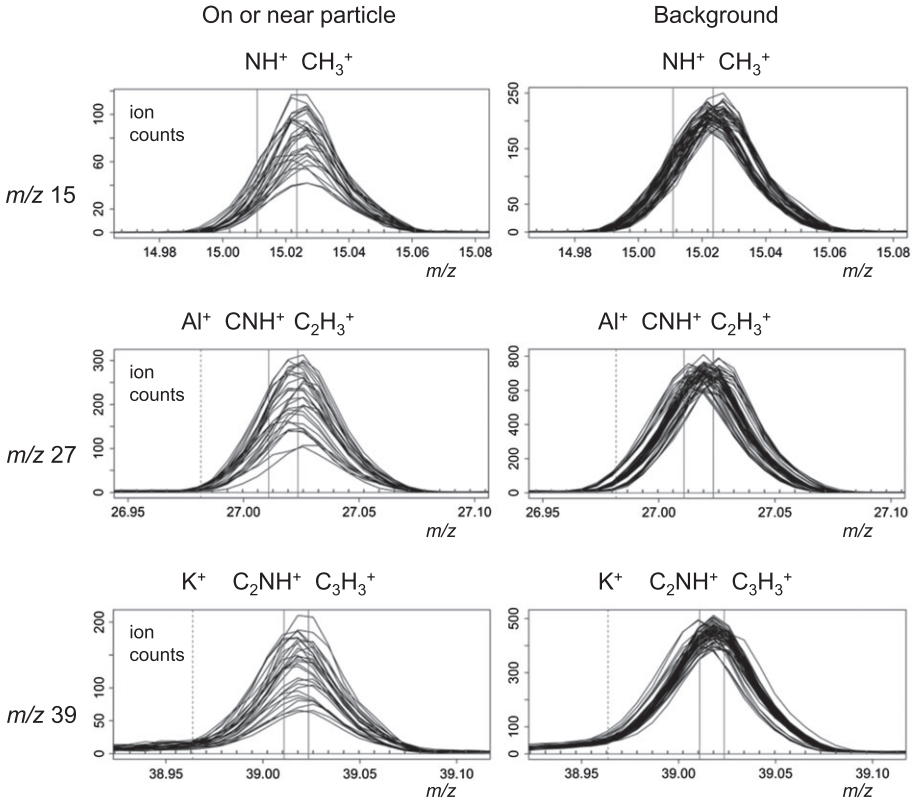


FIGURE 1 Recalibrated mass spectra (examples of peak profiles at masses 15, 27, and 39). Left column: 29 spectra measured on or near the cometary particle *Sai*. Right column: 59 spectra measured on target background. Vertical lines indicate the theoretical masses of elemental and CHNO ions. The small ticks above the mass scale show the mass intervals (bins) corresponding to the variables

not cause relevant changes in the results reported for PCA in Section 4.1 and for the variable importance in Sections 4.2 and 4.3.

2.5.2 | Ion separation and selection of mass intervals for multivariate data

Figure 1 demonstrates the effect of the available mass resolution on the identification of ions. The signals at m/z 15 are unambiguous from CH_3^+, and those at m/z 27 from $C_2H_3^+$ with a possibly small contribution from CNH^+. At m/z 39, the 2 possible organic ions C_2NH^+ and $C_3H_3^+$ cannot be separated, and the corresponding variables reflect a potential mixture of the 2 ions.

We consider all ion formulae with the elements C, H, N, and O in the mass range 12 to 72, fulfilling the chemistry valence rules. Although in mass spectrometry ions with unusual formulae appear, probably not all these formally possible species are sufficiently stable ions. An estimation of the stability, however, was out of the scope of this work. The total number of 288 possible CHNO formulae contains 46 CH ions, 102 CHN ions, 65 CHO ions, and 75 CHNO ions. The degree of unsaturation is between 0 and 7 double-bond equivalents.

For the definition of variables for multivariate data analysis, we select mass intervals in the spectra that cover the masses of the considered ions, assuming Gauss peaks with the means given by the theoretical ion mass m, and the standard deviations s given by m and the mass resolution (Section 2.5.1). Tests showed that mass intervals $m \pm 1.5s$ are a useful compromise between not losing essential information and not including too many waste variables.

Figure 2 shows details for the selection of an appropriate mass range and the corresponding variables around mass 56. Figure 2a contains the ion count signals of 29 spectra from cometary particle *Sai*. The small ticks above the mass scale indicate the mass intervals (bins) used (mass differences approximately 0.01 Da). The theoretical ion masses for $^{56}Fe^+$ and the 9 possible CHNO ions are marked by vertical lines. Figure 2b shows Gauss peaks simulating the intensity profiles of $^{56}Fe^+$ and the CHNO ions (only shown for the lowest and highest masses). The mass range selected reaches from the lowest to highest masses of the CHNO ions, extended by $\pm 1.5s$. This mass range contains 16 mass intervals (bins) ranging from 55.95544 to 56.09710 Da; the ion counts at these m/z values are used as variables for data analysis. Note that the lowest selected mass bins may be affected by Fe ions. Assuming a complex mixture of organic ions and considering the presence of all formally possible CHNO ions, a deconvolution of all CHNO ion signals is not feasible. The ion counts at the mass bins (corresponding to the variables) can be considered as linear combinations of the signals from typically 2 to 15 CHNO ions.

In addition to the described formal selection of relevant mass bins, a visual inspection of the spectra showed that the mass numbers 16, 17, 20, 33 to 36, 48, and 49 contain only very weak signals for CHNO ions with typically less than 3

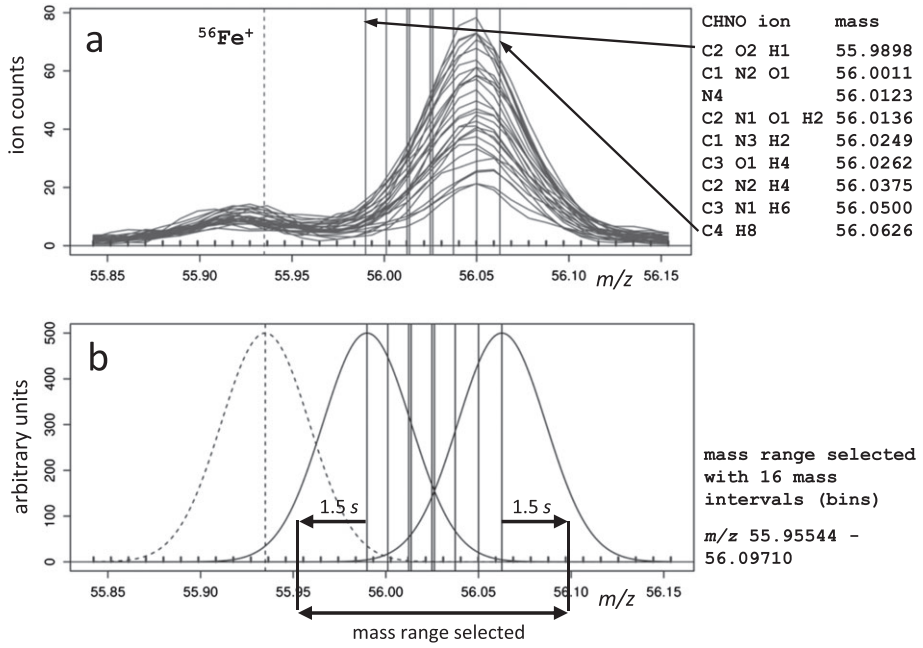


FIGURE 2 Selection of an appropriate mass range and the corresponding mass bins (variables) for the recognition of CHNO ions around mass 56

counts per mass bin. These mass bins and the corresponding ion formulae have been excluded from further evaluations. The peak from $^{28}Si^+$ (from inorganic cometary material and from the contamination PDMS) is usually high, and the peak tail reaches potential signals from CHNO ions. Thus, also bins and formulae around mass 28 have been eliminated. In total, 276 CHNO ions are considered, ranging from $^{12}C^+$ to $^{12}C_5H_{12}{}^+$, covering 540 mass bins (out of 8122 in the mass range 11.5 to 72.5 Da), which define the $M = 540$ variables used for data analysis. The original variables contain the number of ions counted in these mass bins. Because only relative ion counts are relevant here, the variables were normalized to a constant sum in each spectrum. Normalization to a constant sum of secondary ions reflects the competing processes for ion formation and fragmentation that characterize the chemical composition of the sample. Some methods use specific data transformations considering the compositional type of the data (Section 3.4).

3 | DATA ANALYSIS METHODS

The used data consist of the matrices $\boldsymbol{X}_1$ (29 × 540) and $\boldsymbol{X}_2$ (23 × 540) for spectra measured on the cometary particles *Sai* and *Kerttu*, respectively. Matrix $\boldsymbol{X}_0$ (59 × 540) is from spectra measured on the background. Matrix element x_{ij} is the number of secondary ions counted in spectrum i and mass bin j. The data structure—especially the separation of particles versus background—has been visualized by PCA (Section 4.1).

The variables are related to ion masses and consequently to 1 or several ion formulae CHNO. In a univariate approach, each variable is considered separately, eg, by applying the t test, resulting in a probability p for the zero hypothesis H_0 "data of both classes are from the same population." For multivariate discrimination methods, 3 different criteria for the importance of variables for discrimination have been used to search for CHNO ions that may be relevant for the cometary material. In the next sections, we define the used methods.

3.1 | Statistical *t* and *u* tests

For a comparison of the class centers for each variable separately, the t test and the u test (Mann-Whitney-Wilcoxon test) were applied; data were normalized to a constant sum for the objects. The criterion for variable importance used is

$$\mathrm{LOG}p = \mathrm{sgn}[-_{10}\log(p)], \tag{1}$$

with p for the probability of H_0; sgn = +1 if the central value (the mean or alternatively the median) of the particle class is higher than that for the background class, else sgn = −1. LOGp has a high value if the mean of the particle class is significantly higher than that for the background class. An advantage of this univariate approach is the evident interpretation in terms of a single, potentially characteristic ion.

3.2 | Discriminant partial least squares

Discriminant partial least squares classification (D-PLS) creates from given multivariate data (centered $\boldsymbol{X}$) a linear discriminant variable, defined by a vector $\boldsymbol{b}_{\mathrm{PLS}}$ containing regression coefficients for the variables.[23] Standardized regression coefficients are scaled by the standard deviation of the corresponding variable j (b_j/s_j) and are suitable to characterize the variable importance. The x data were normalized to a constant sum for the objects. An object i is assigned to the particle class if $\mathbf{x}_i{}^T \bullet \boldsymbol{b}_{\mathrm{PLS}} > 0$; otherwise, it is assigned to the background class. The optimum number of PLS components (A_{opt}) has been estimated by the repeated double cross-validation strategy.[24,25] Parameters for repeated double cross-validation were as follows: 4 segments in the outer loop (split into a calibration set and a test set), 5 segments in the inner loop (estimation of the optimum number of components for the actual calibration set), and 50 repetitions. Among the 4 × 50 obtained estimations of A_{opt}, the value 6 had the highest frequency and was used for a final D-PLS model from all objects, giving the regression coefficients $\boldsymbol{b}_{\mathrm{PLS}}$ for the M variables. This linear, multivariate approach is capable of finding groups of variables that are together responsible for a class separation—even if each single variable may have only poor discrimination power.

3.3 | Random forest

Random forest (RF) classification is a nonlinear method based on decision trees that use a subset of randomly selected variables to partition the variable space, based on classification and regression trees (CART).[26] An ensemble of decision

trees (the forest) is evaluated for the purpose of classification or regression.[27] The training is performed with random samples of the objects. The variable importance for class separation can be estimated by the *mean decreasing accuracy* (MDA), as implemented in the R function *randomForest* (package *randomForest*[28]). Mean decreasing accuracy has a high value if the classification accuracy decreases considerably when the variable is eliminated. A variable with a high MDA is important relative to the others. Random forest was applied with the x data normalized to a constant sum for the objects and with 500 trees per function call. The final MDA criterion is the arithmetic mean of 50 repetitions.

3.4 | Robust pairwise log ratios

The method rPLR has been developed for the identification of biomarkers using binary classification.[29] The method rPLR is independent from the absolute values of the M variables because it uses the log ratios of all pairs (j, k) of the variables; for object i given by $\ln(x_{ij}/x_{ik})$ with $j, k = 1, ..., M$. A variation matrix for all objects, $\boldsymbol{T}(M \times M)$, is defined by the matrix elements

$$t_{jk} = \text{var}\left[\ \ln\left(x_{1j}/x_{1k}\right), ..., \ln\left(x_{Nj}/x_{Nk}\right)\ \right], \tag{2}$$

$j, k = 1, ..., M$; N is the number of all objects; var denotes the robust τ estimator of variance.[30] $\boldsymbol{T}$ is symmetric, and the diagonal elements are 0. The matrix elements characterize the variabilities of the log ratios of all pairs of variables. Analogously, the variation matrices ${}_1\boldsymbol{T}$ and ${}_2\boldsymbol{T}$ are defined by solely using the N_1 and N_2 objects of the particle class and the background class (matrix elements are ${}_1t_{jk}$ and ${}_2t_{jk}$), respectively. A variable j with high importance for classification (a biomarker) will have considerably different elements ${}_1t_{jk}$ and ${}_2t_{jk}$ and tentatively smaller than t_{jk} in $\boldsymbol{T}$ (for all k). A statistic V_j for estimation of the importance of variable j has been proposed[29] as

$$V_j = \text{sum}\left\{\left[N_1\,{}_1t_{jk}{}^{0.5} + N_2\,{}_2t_{jk}{}^{0.5}\right]/Nt_{jk}{}^{0.5}\right\} \qquad \text{sum for } k = 1, ..., M. \tag{3}$$

V_j is approximately normally distributed, as well as the normalized version suggested for practical use:

$$V_j^* = -\left[V_j - mean(V_k)\right]/sd(V_k), \tag{4}$$

with *mean* for the arithmetic mean, *sd* for the standard deviation, and $k = 1, ..., M$. Big values for V^* indicate a high variable importance (owing to the minus sign in the above equation); a reasonable cutoff is the 0.975 quantile. V_j^* refers to a single variable j; however, all bivariate relationships between the other variables and variable j are considered. This criterion for variable importance is complementary to the approaches based on D-PLS or RF.

We applied rPLR as implemented in the R function *biomarker* (library *robCompositions*). A practical limitation of rPLR is the requirement for positive nonzero values of the original variables; a solution for this problem—based on a defined detection limit—was described.[31] Here, a simple procedure was performed as follows: Ion counts equal to 0 are replaced by uniformly distributed noise between 0.1 and 0.2 (the median of all ion counts is 33.4, and the maximum is 1907).

4 | RESULTS

4.1 | Exploration

For an insight into the structure of the data sets from 2 cometary particles and the background, PCA score and loading plots have been used. Figure 3 shows the results obtained from a subset of the variables covering the mass range 12 to 15, which has been shown to be characteristic for the carbonaceous substances in the dust of the comet 67P.[16] The number of variables (mass bins) was $M = 21$ (4 for m/z 12, 4 for m/z 13, 6 for m/z 14, and 7 for m/z 15), and the number of objects (spectra) is $N = 111$ (29 for particle *Sai*, 23 for *Kerttu*, and 59 for the background). The variables have been normalized to a constant sum of 100.

The first 2 PCA components preserve almost 95% of the total variance; the background spectra are clearly separated from the particle spectra, which exhibit some overlap of the classes *Sai* and *Kerttu*, while the background spectra are divided into 2 clusters. The loading plot indicates relatively high values for the variables from m/z 12 (C^+) for particle spectra and for the variables from m/z 15 (CH_3^+) for background spectra. The variation in the second PCA component

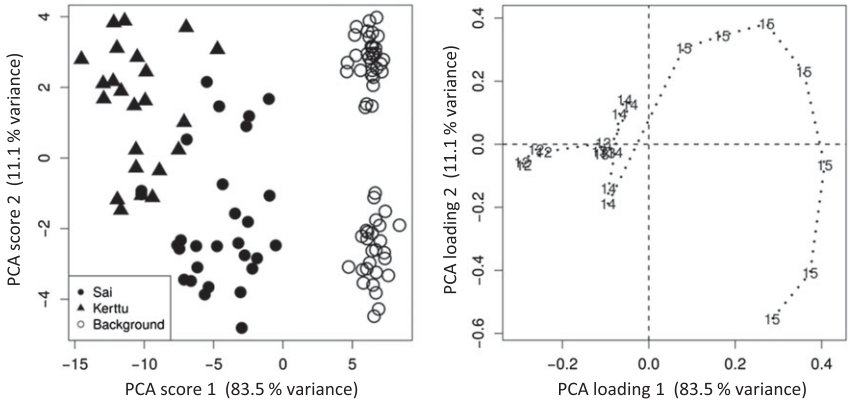


FIGURE 3 Principal component analysis (PCA) score and loading plot for spectral data measured on the cometary particles *Sai* and *Kerttu* and on the background. The number of objects is 111; the 21 variables (mass bins) are from the mass range 12-15 Da and have been normalized to a constant sum. The dotted lines in the loading plot connect variables with adjacent masses; the variables are denoted by the integer mass

may be due to minor contributions of N^+ (m/z 14) and NH^+ (m/z 15); however, an artifact of the peak shape cannot be excluded.

As discussed in Section 2.5.1, the mass scale has errors of typical ±1 bins that may cause wrong assignments of the variables by ±1 indices. Note that the signals of a specific ion in the mass range 12 to 15 are distributed among about 12 bins. The influence of these uncertainties on PCA results has been estimated by a simulation of errors caused by a random exchange of values in neighboring columns of $\boldsymbol{X}$. A data set with partial randomization in 10 randomly selected neighboring columns and in 162 randomly selected rows (30% of total 540 rows) gave a PCA score plot with very similar clustering as obtained from the original data; the variances preserved in the first 2 components were slightly reduced to 70.4% and 8.6% of the total variance.

4.2 | Univariate variable importance

The t test and the u test have been applied to each variable for estimating their importance for a discrimination between the cometary particle data and the background data. The variables have been normalized to a constant sum of 100. Figures 4 and 5 show the results from the t test for the mass range 12 to 15 for *Sai* and *Kerttu*,

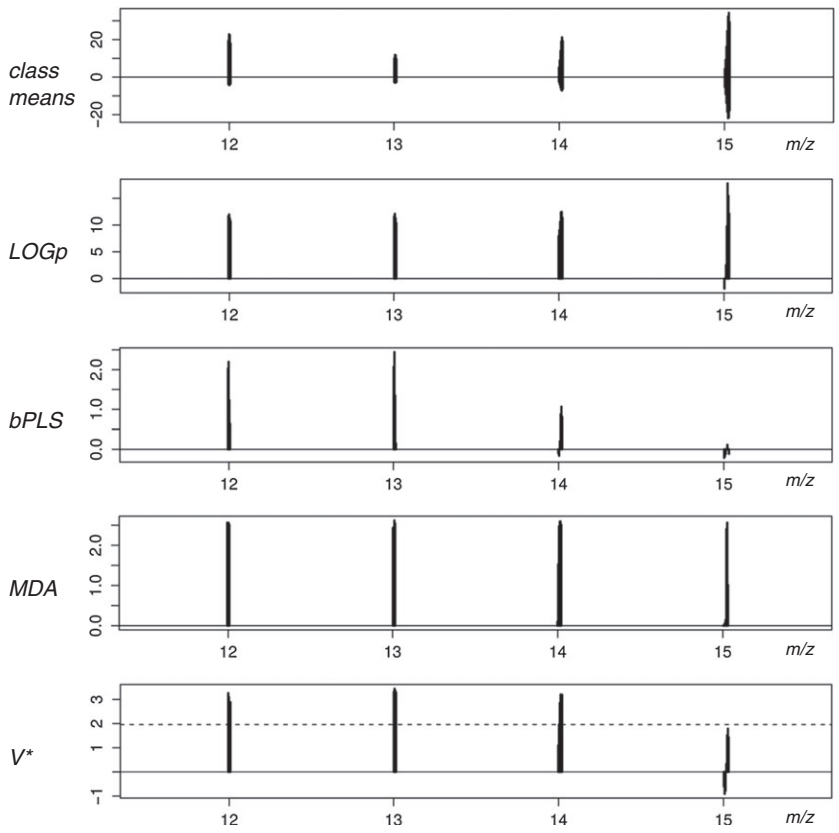


FIGURE 4 Significance of variables for discrimination between spectra measured on cometary particle *Sai* and on background in mass range 12 to 15 Da

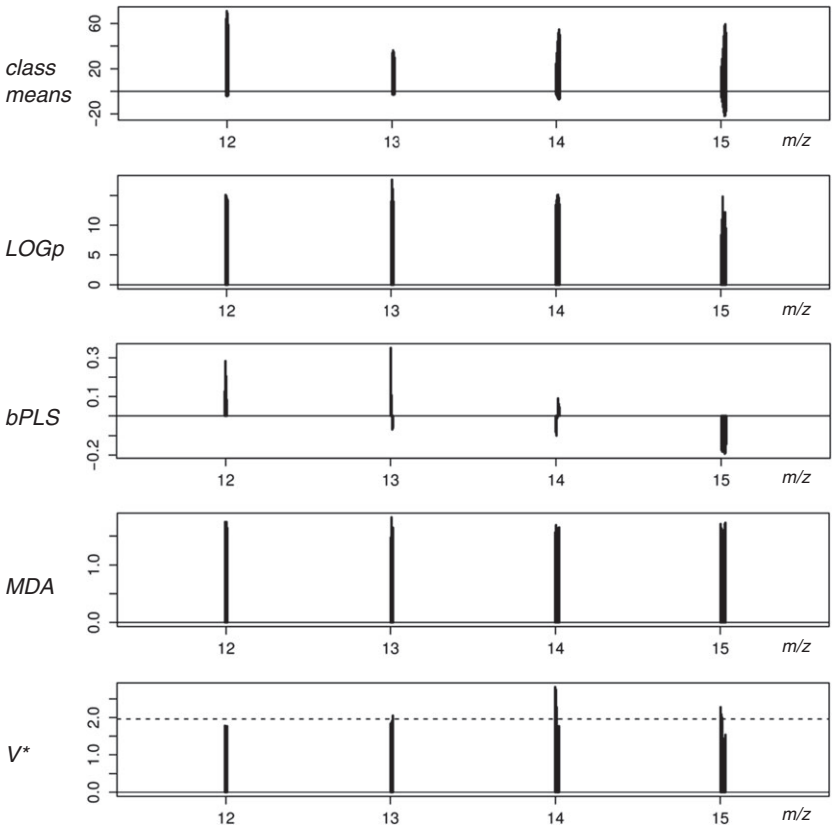


FIGURE 5 Significance of variables for discrimination between spectra measured on cometary particle *Kerttu* and on background in mass range 12 to 15 Da

respectively. The first plot in the figure compares the class means with positive bars for the cometary particle class, and negative bars for the background class (unit of the ordinate is 10^4 x/sum(x_j), $j = 1, ..., M$). The second plot shows the result LOGp from the t test as defined in Equation 1. Except a few mass bins at m/z 15, all others show significantly higher means for the particle spectra than for the background spectra. These results clearly indicate the presence of carbonaceous compounds in the investigated cometary matter, as previously stated on the basis of ion ratio evaluations.[16]

Results from the u test are not shown because they are almost identical to the results from the t test. The Pearson correlation coefficients between the LOGp values of the t and the u test are 0.932 and 0.922 for the *Sai* and the *Kerttu* data, respectively.

4.3 | Multivariate variable importance

Plots 3 to 5 in Figures 4 and 5 show the variable importance criteria obtained by D-PLS (criterion standardized PLS regression coefficient bPLS), RF classification (criterion MDA), and the method based on rPLR (criterion V^*). The multivariate methods have been applied to the complete data sets with 540 variables.

The results for MDA and V^* show for all variables a high importance for the separation of the particle class and the background class; V^* is mostly above or near the 0.975 quantile (dashed line). The data sets for particles *Sai* and *Kerttu* give similar results, thus confirming the outcome of the t test.

The standardized regression coefficients bPLS of the discriminant variable show similar results as the other methods, and the positive signs point to high values of the corresponding ion counts in the particle class. The regression coefficients of variables around m/z 15 (CH_3^+) are negative, corresponding to the PCA loading plot and confirming the higher relevance of the ions C^+, CH^+, and CH_2^+ for cometary material.[16]

The results for the variable importance obtained from multivariate approaches reflect the influence of all variables together, and an interpretation of single variables may be difficult, especially if the number of variables is large.[32] A partial randomization of $\boldsymbol{X}$ by exchanging randomly selected matrix elements in neighboring columns (as described in Section 4.1) had only a minor influence on the variable importance.

4.4 | Ion formulae

The potentials and limitations of recognizing CHNO ions from secondary ion mass spectral signals measured with the COSIMA instrument are discussed for the mass regions at m/z 39 and 55 (Figure 6). The shown variable importance measures have been estimated from the data set for the *Sai* particle; the first row in the figure contains the original ion count data for the 29 spectra measured on this particle.

At m/z 39, only the 2 C-containing ions C_2HN^+ (39.0109 Da) and $C_3H_3^+$ (39.0235 Da) are possible; ions $^{39}K^+$ (38.9637 Da) do not appear in the signals and would be well separated from the organic ions. The variable importance measures LOGp and bPLS indicate an enhanced presence of $C_3H_3^+$ in the cometary matter. The criteria MDA and V^* show somewhat bigger values for the mass bins near the mass of this ion than for the other one. We conclude from the data that $C_3H_3^+$ ions are mostly from the surface of the cometary particle.

At m/z 55, 6 CHNO ions are formally possible: C_2NOH^+ (55.0058 Da), CN_3H^+ (55.0171 Da), $C_3OH_3^+$ (55.0184 Da), $C_2N_2H_3^+$ (55.0296 Da), $C_3NH_5^+$ (55.0422 Da), and $C_4H_7^+$ (55.0548 Da). Elemental ion $^{55}Mn^+$ (54.9381 Da) does not appear and would be well separated from the organic ions. Even if some of the listed CHNO ions are excluded, the mass resolution of about 1000 is not sufficient to establish the presence and/or absence of the CHNO ions unambiguously. Furthermore, from the contamination PDMS, ions, $C_2H_3Si^+$ (55.0004 Da, dotted line) may be produced. The broad spectral peak at mass 55.05 is probably mainly from $C_4H_7^+$; the LOGp and bPLS criteria indicate a background origin. The criteria MDA and V^* do not give a clear picture. We conclude that that $C_4H_7^+$ and perhaps $C_3NH_5^+$ are probably from the background.

We summarize the results as follows (ions not separable by their mass are denoted in brackets, eg, $\{CH_2^+, N^+\}$).

1. Secondary ions probably from cometary material are mainly unsaturated species: C^+; CH^+; $\{CH_2^+, N^+\}$; $\{CH_3^+, NH^+\}$; $C_2H_2^+$; $\{C_2H_3^+, CNH^+\}$; C_3^+; C_3H^+; $\{C_3H_2^+, C_2N^+\}$; $\{CH_2CN^+, C_3H_4^+\}$; $C_3H_3^+$; and C_4^+. The presence of N-containing ions cannot be excluded from the spectral data or the variable importance data; however, a clear evidence for the presence of these ions could not be derived. Note that from negative secondary ion mass spectrometry data, only a small amount of N-containing substances was recognized, with about 3.5% of the C atoms being N atoms.[15]

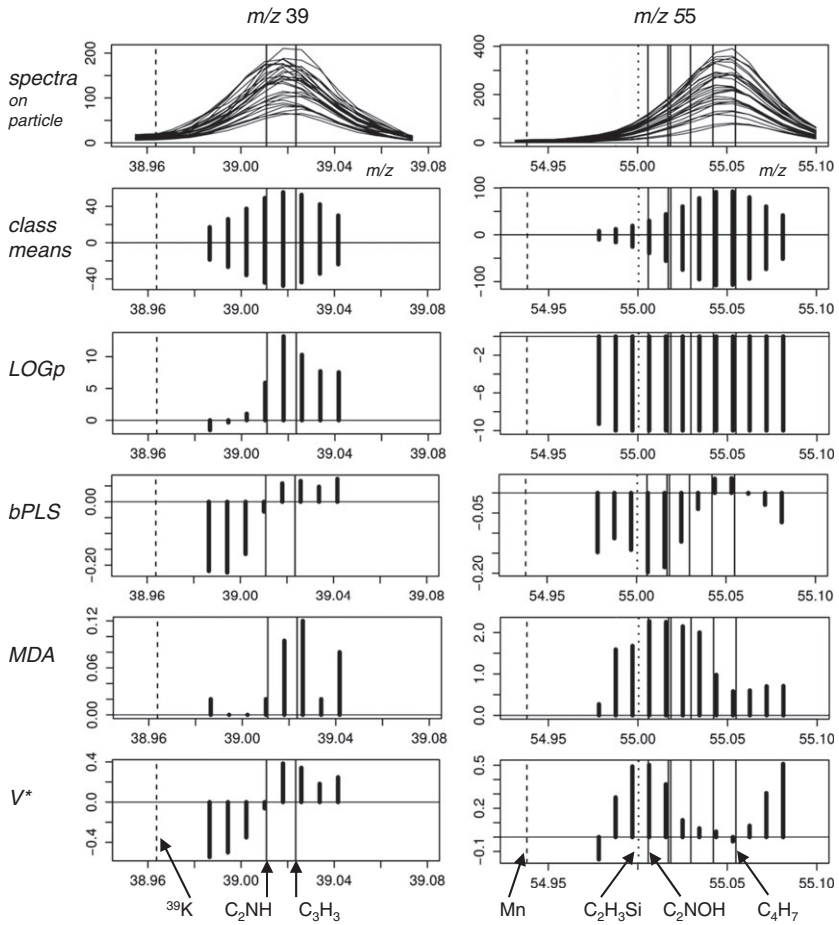


FIGURE 6 Significance of variables for discrimination between spectra measured on the cometary particle *Sai* and on background

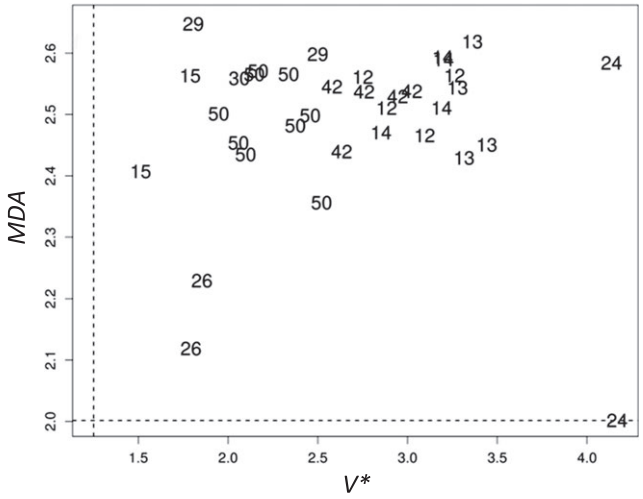


FIGURE 7 Variables with importance criteria V^* and mean decreasing accuracy (MDA) above the 0.9 quantiles (dashed lines). The variables are denoted by the integer masses of the corresponding mass bins; results are from the discrimination between spectra from the cometary particle *Sai* and the background

2. Secondary ions probably from the background are mainly saturated or almost saturated ions, made of C and H atoms, eg, $C_3H_{5\text{-}9}^+$, $C_4H_{7\text{-}9}^+$, and $C_5H_{7\text{-}12}^+$. The sources are probably the contamination by PDMS and other substances with aliphatic-like chains of unknown origin.[33]

Figure 7 shows the variables (mass bins) with the criterions MDA and V^* both above the 0.9 quantiles, which therefore can be considered as important variables for a discrimination of spectra from particle *Sai* and the background. The 34 variables found are from 10 mass numbers. Corresponding to Figure 4, variables for m/z 12 to 15 are present in this group (ions $CH_{0\text{-}3}^+$). Mass 24 is for C_2^+ and mass 26 for $C_2H_2^+$, both with higher relative abundances on the particle than on the background. At mass number 29, the found variables are related to the masses 28.9885 and 28.9953, denoting the ions SiH^+ (mass 28.9848) and COH^+ (mass 29.0027) but not $C_2H_5^+$ (mass 29.0391). A similar situation appears at mass 30, indicating the ions COH_2^+ and SiH_2^+. The Si-containing ions are at least partially from the PDMS contamination. For mass 42, several CHNO ions are formally possible, highly probable is $C_3H_6^+$ with background origin. No reasonable interpretation for the found variables at mass 50 can be given because the spectra show in this mass range a series of not resolved peaks with low intensity.

5 | CONCLUSIONS

The application of 4 complementary methods for estimating the variable importance in binary classification gives a more complete picture than a single method. The conclusions are based on results obtained from data sets with 540 variables and 23 to 59 objects (spectra) per class. The results of the univariate *t* test are directly interpretable in terms of the meaning of the variables (here ion formulae). The *u* test showed almost identical results as the *t* test. The standardized regression coefficients for a linear discriminant variable—calculated by D-PLS—provided similar results as the *t* test, although based on a multivariate concept. The other 2 multivariate classification methods applied are nonlinear and robust; the RF is based on decision trees, and the rPLR method on the variances of logarithms of all possible ratios of the variables. The variable importance criteria of these 2 methods indicate groups of variables that are relevant for the class discrimination.

The evaluation of secondary ion mass spectrometry data measured at cometary particles proved the presence of carbonaceous substances. No distinct organic substance classes are evident from the data; however, a complex mixture of unsaturated organic compounds may be present. The results are consistent with the previously claimed presence of high–molecular weight structures.[13,16] The applied criteria for univariate and multivariate variable importance for discriminating spectra from cometary particles and from background indicate clearly various secondary ions from cometary material, such as $CH_{0\text{-}3}^+$, $C_2H_{0\text{-}3}^+$, $C_3H_{0\text{-}4}^+$, and C_4^+. The presence of other CHNO ions, eg, C_2N^+, CH_2CN^+, and COH^+, is possible owing to the results from *t* tests and D-PLS.

ACKNOWLEDGEMENTS

This work was supported by the Austrian Science Fund (FWF), project P 26871-N20. J. W. thanks the FWF and the Czech Science Fund (GACR), project I 1910-N26, for their support.

The COSIMA instrument was built by a consortium led by the Max-Planck-Institut für Extraterrestrische Physik, Garching, Germany, in collaboration with the Laboratoire de Physique et Chimie de l'Environnement et de l'Espace, Orléans, France; the Institut d'Astrophysique Spatiale, CNRS/Université Paris Sud, Orsay, France; the Finnish Meteorological Institute, Helsinki, Finland; the Universität Wuppertal, Wuppertal, Germany; von Hoerner und Sulger GmbH, Schwetzingen, Germany; the Universität der Bundeswehr, Neubiberg, Germany; the Institut für Physik, Forschungszentrum Seibersdorf, Seibersdorf, Austria; and the Institut für Weltraumforschung, Österreichische Akademie der Wissenschaften, Graz, Austria, and is led by the Max-Planck-Institut für Sonnensystemforschung, Göttingen, Germany. The support of the national funding agencies of Germany (DLR, grant 50QP1302), France (CNES), Austria, Finland, and the ESA Technical Directorate is gratefully acknowledged. The authors thank the other members of the COSIMA team for their contributions.

ORCID

Kurt Varmuza http://orcid.org/0000-0002-3534-4001

REFERENCES

1. Wikipedia, 67P/Churyumov–Gerasimenko. https://en.wikipedia.org/wiki/67P/Churyumov%E2%80%93Gerasimenko. Accessed Nov 28, 2017.
2. ESA, Rosetta. http://www.esa.int/Our_Activities/Space_Science/Rosetta/The_Rosetta_orbiter. Accessed Nov 28, 2017.
3. Schulz R, Boehnhardt AC, Glassmeier KH (Eds). *Rosetta: ESA's Mission to the Origin of the Solar System*. New York: Springer; 2009.
4. Wikipedia, Rosetta (spacecraft). https://en.wikipedia.org/wiki/Rosetta_(spacecraft)#Background. Accessed Nov 28, 2017.
5. Altwegg K, Balsiger H, Berthelier JJ, et al. Organics in comet 67P—a first comparative analysis of mass spectra from ROSINA-DFM, COSAC and Ptolemy. *MNRAS (Mon Not Roy Astron Soc)*. 2017;479:S130-S141.
6. Kissel J, Altwegg K, Clark BC, et al. COSIMA—high resolution time-of-flight secondary ion mass spectrometer for the analysis of cometary dust particles onboard Rosetta. *Space Sci Rev*. 2007;128(1-4):823-867.
7. Langevin Y, Hilchenbach M, Ligier N, et al. Typology of dust particles collected by the COSIMA mass spectrometer in the inner coma of 67P/Churyumov Gerasimenko. *Icarus*. 2016;271:76-97.
8. Merouane S, Stenzel O, Hilchenbach M, et al. Evolution of the physical properties of dust and cometary dust activity from 67P/Churyumov-Gerasimenko measured in situ by Rosetta/COSIMA. *MNRAS (Mon Not Roy Astron Soc)*. 2017;469(Suppl_2):S459-S474.
9. Merouane S, Zaprudin B, Stenzel O, et al. Dust particle flux and size distribution in the coma of 67P/Churyumov-Gerasimenko measured in situ by the COSIMA instrument on board Rosetta. *Astron Astrophys*. 2016;596:A87.
10. Hornung K, Merouane S, Hilchenbach M, et al. A first assessment of the strength of cometary particles collected in-situ by the COSIMA instrument onboard ROSETTA. *Planet Space Sci*. 2016;133:63-75.
11. Brownlee D. The Stardust mission: analyzing samples from the edge of the solar system. *Annu Rev Earth Planet Sci*. 2014;42(1):179-205.
12. Stephan T, Flynn GJ, Sandford SA, Zolensky ME. TOF-SIMS analysis of cometary particles extracted from Stardust aerogel. *Meteoritics Planet Sci*. 2008;43(1-2):285-298.
13. Bardyn A, Baklouti D, Cottin H, et al. Carbon-rich dust in comet 67P/Churyumov-Gerasimenko measured by COSIMA/Rosetta. *MNRAS (Mon Not Roy Astron Soc)*. 2017;469(Suppl_2):S712-S722.
14. Stenzel O, Hilchenbach M, Merouane S, et al. Similarities in element content between comet 67P/Churyumov-Gerasimenko coma dust and selected meteorite samples. *MNRAS (Mon Not Roy Astron Soc)*. 2017;469(Suppl_2):S492-S505.
15. Fray N, Bardyn A, Cottin H, et al. Nitrogen-to-carbon atomic ratio measured by COSIMA in the particles of comet 67P/Churyumov-Gerasimenko. *MNRAS (Mon Not Roy Astron Soc)*. 2017;469(Suppl_2):S506-S516.
16. Fray N, Bardyn A, Cottin H, et al. High-molecular-weight organic matter in the particles of comet 67P/Churyumov-Gerasimenko. *Nature*. 2016;528:72-74.
17. Hornung K, Kissel J, Fischer H, et al. Collecting cometary dust particles on metal blacks with the COSIMA instrument onboard ROSETTA. *Planet Space Sci*. 2014;103:309-317.
18. Xu Y, Brereton RG. Diagnostic pattern recognition on gene-expression profile data by using one-class classification. *J Chem Inf Model*. 2005;45(5):1392-1401.

19. Hubert M, Rousseeuw PJ, Vanden Branden K. ROBPCA: a new approach to robust principal components. *Technometrics*. 2005;47(1):64-79.
20. Varmuza K, Filzmoser P, Hilchenbach M, et al. Selected chemometric approaches for mass spectra from comet dust grains (Rosetta). 14th Scandinavian Symposium on Chemometrics, 14-17 June 2015, Chia, Sardinia, Italy 2015: Poster presentation.
21. Paquette J, Engrand C, Stenzel O, Hilchenbach M, Kissel J. Searching for calcium-aluminum-rich inclusions in cometary particles with Rosetta/COSIMA. *Meteoritics Planet Sci*. 2016;51(7):1340-1352.
22. R. A language and environment for statistical computing. Vienna, Austria: R Development Core Team, Foundation for Statistical Computing, www.r-project.org; 2017.
23. Brereton RG, Lloyd GR. Partial least squares discriminant analysis: taking the magic away. *J Chemometr*. 2013;28:213-225.
24. Filzmoser P, Liebmann B, Varmuza K. Repeated double cross validation. *J Chemometr*. 2009;23(4):160-171.
25. Varmuza K, Filzmoser P. Repeated double cross validation (rdCV)—a strategy for optimizing empirical multivariate models, and for comparing their prediction performances. In: Khanmohammadi M, ed. *Current applications of chemometrics*. New York, NY, USA: Nova Science Publishers; 2015:15-31.
26. Breiman L, Friedman J, Stone CJ, Olshen RA. *Classification and Regression Trees*. New York, USA: Chapman & Hall; 1984.
27. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5-32.
28. Liaw A, Wiener M. Classification and regression by random forest. *R News*. 2002;2:18-22.
29. Walach J, Filzmoser P, Hron K, Walczak B, Najdekr L. Robust biomarker identification in a two-class problem based on pairwise log-ratios. *Chemom Intell Lab Syst*. 2017;171:277-285.
30. Yohai VJ, Zamar RH. High breakdown-point estimates of regression by means of the minimization of an efficient scale. *J Am Stat Assoc*. 1988;83(402):406-413.
31. Templ M, Hron K, Filzmoser P, Gardlo A. Imputation of rounded zeros for high-dimensional compositional data. *Chemom Intell Lab Syst*. 2016;155:183-190.
32. Wehrens R, Franceschi P, Vrhovsek U, Mattivi F. Stability-based biomarker selection. *Anal. Chim. Acta*. 2011;705(1-2):15-23.
33. Hilchenbach M, Kissel J, Langevin Y, et al. Comet 67P/Churyumov-Gerasimenko: close-up on dust particle fragments. *Astrophys J Lett*. 2016;816(2):L32

How to cite this article: Varmuza K, Filzmoser P, Hoffmann I, et al. Significance of variables for discrimination: Applied to the search of organic ions in mass spectra measured on cometary particles. *Journal of Chemometrics*. 2018;e3001. https://doi.org/10.1002/cem.3001