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], Johan Silén[4], Oliver Stenzel[5], Jochen Kissel[5], Martin Hilchenbach[5]

[1] Vienna University of Technology, Institute of Statistics and Mathematical Methods in Economics, Research Unit Computational Statistics, Vienna, **Austria**
[2] Université Paris-Est Créteil et Université Paris Diderot, Créteil, **France**; [3] Université d'Orléans, Orléans, **France**
[4] Finnish Meteorological Institute, Helsinki, **Finland**; [5] Max Planck Institute for Solar System Research (MPS), Göttingen, **Germany**

AIM. Recognition of mass spectral signals from CHNO (organic) compounds present at the surface of particles that have been collected near a comet (ESA mission Rosetta, TOF-SIMS instrument).

STRATEGY. Multivariate discrimination of spectra measured on a particle and on the background. Significance of variables indicates potentially relevant CHNO ions originating from the particle.

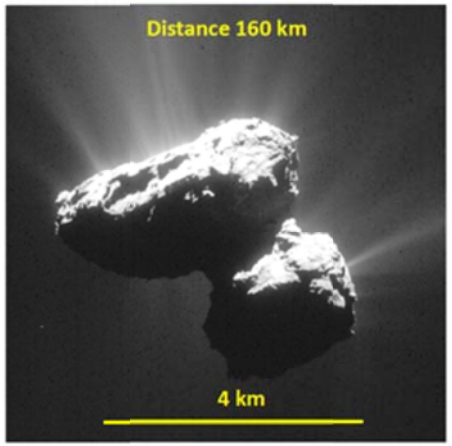
Poster at CC 2017 – *Conferentia Chemometrica,* 3 – 6 Sep 2017, Gyöngyös-Farkasmály, Hungary

http://cc2017.ttk.mta.hu/

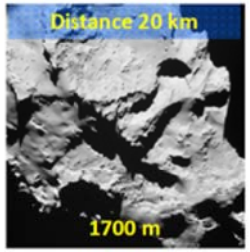
* kurt.varmuza@tuwien.ac.at; www.lcm.tuwien.ac.at/vk/

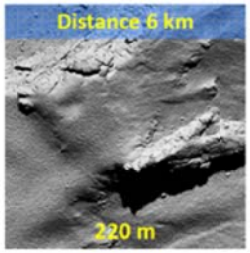
Version 170901

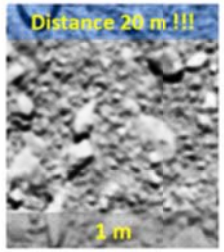
The Comet [1]



Name: **67P/Churyumov-Gerasimenko**
Orbit: 1.2 – 5.7 AU from sun, 6.4 years.
Rotation: 12.76 h. Density: 0.4-0.5 g/cm^3.
Albedo: 5 % reflectance (very black).





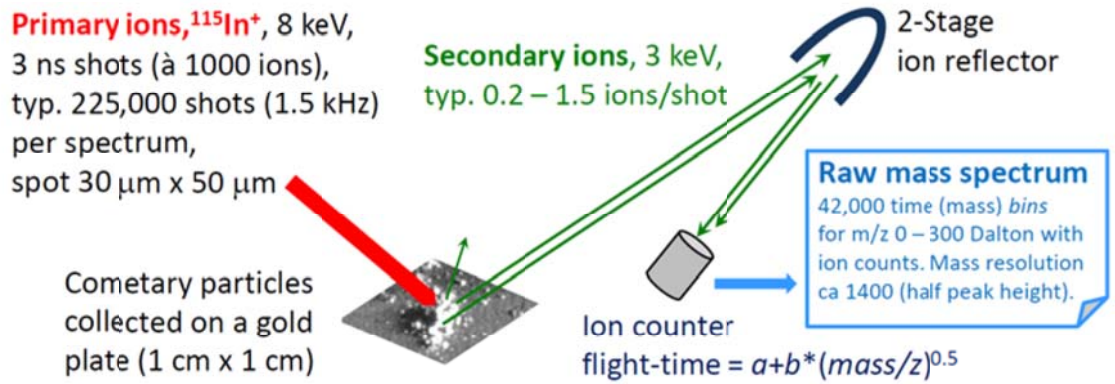


The Mission (Rosetta) [2]

This mission of the European Space Agency (ESA) is named after the ***Rosetta Stone*** (196 BC), bearing an inscription that was the key to the decipherment of Egyptian hieroglyphs.

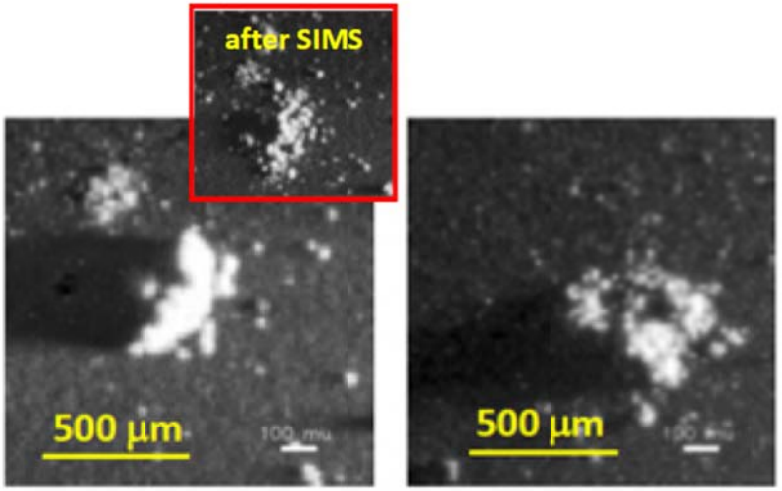
2 Mar 2004 Launch.
6 Aug 2014 Arrival (100 km).
Escorting the comet; typ. distance 10 – 200 km; 1.2 – 3.8 AU from sun.
12 Nov 2014 *Philae* landing.
30 Sep 2016 Controlled landing of Rosetta on comet (2 m/s), switch-off.

The Instrument (COSIMA) [3]



Samples

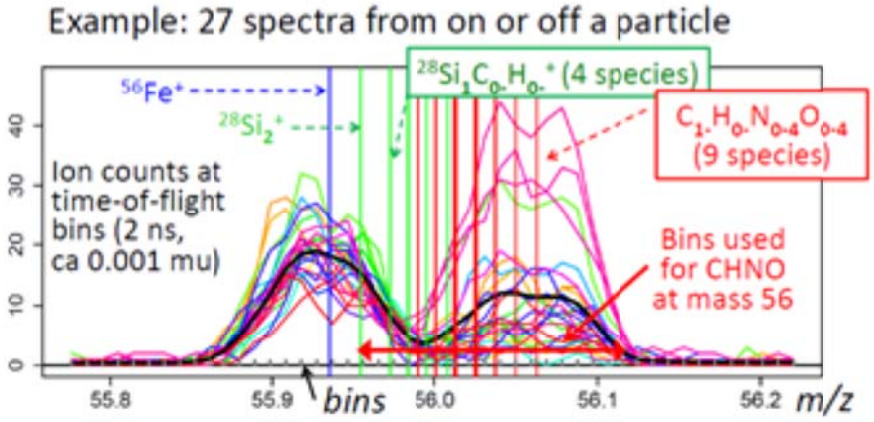
On metal targets (mostly Au) 1 cm x 1 cm more than 35,000 cometary particles were collected and imaged. Size 10 - 1000 µm. Fluffy material; some particles compact, others disrupted before collection [4].



Spectral Data

For about 200 particles more than 30,000 SIMS spectra were measured [5].

Preprocessing: Mass recalibration and rebinning; selection of mass bins (variables) for the considered CHNO ions.



Chemical Formulae of Ions

Considered mass range: 12 – 72 *

No. of formulae C_{0-} H_{0-} N_{0-4} O_{0-4}: 322

No. of mass bins (variables): 665 **

* Avoiding ions $Si(CH_3)_3^+$, mass 73.047, from contamination PDMS (polydimethylsiloxane).

** Mass resolution 1000, ± 1.5 *sd* (Gaussian peaks).

Formulae	*No.*
C H	50
C H N	112
C H O	75
C H N O	85
Total	322

Methods

Following methods have been used to characterize the importance of variables for a discrimination between comet particle spectra (class 1) and background spectra (class 2).

A high variable importance indicates (via the mass corresponding to the variable) one or several ion species that may be prominent for the particle or for the background.

- **t-test** for comparing the class means (univariate; u-test gives almost identical results)

  Data preprocessing: Normalization to constant sum of variables

  Criterion: ***LOGp*** = *sgn* [-log(*p*)] *p*, probability of H_0; *sgn* = +1 if $mean_{CLASS1} > mean_{CLASS2}$, else *sgn* = -1

- **D-PLS**

  Data preprocessing: Normalization to constant sum of variables

  D-PLS: Optimum no. of PLS components estimated by repeated double cross validation (rdCV) [6]

  Criterion: $\boldsymbol{b_{PLS}}$ = standardized regression coefficients of discriminant variable (>0 indicates class 1)

■ Random Forest (RF) classification

Data preprocessing: Normalization to constant sum of variables

RF: R library *randomForest*, function *randomForest*

Criterion: ***MDA*** (Mean Decreasing Accuracy, mean of 50 repetitions); a variable is important for class discrimination if the classification accuracy decreases considerably when the variable is eliminated.

■ Robust Pair-wise Log-Ratios (rPLR)

Data preprocessing: Values $x_{ij} = 0$ are replaced by uniformly distributed noise between 0.1 and 0.2 (median of all x is 33.4, maximum is 1907).

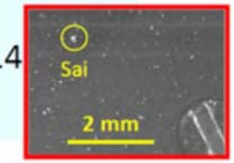
rPLR: R library *robCompositions*, function *biomarker* [7]

Criterion: ***V**** is a newly developed statistics [7], derived from three variation matrices (for all objects, and for objects of the classes separately) containing robust estimations of the variances of all variable pairs (log-ratio transformed). The distribution of V^* is approximately standard normal, and quantile 0.975 is used to indicate the importance of the variables (biomarkers). The variable importance of a variable considers all other variables, and is not affected by the size effect, as ratios of variables are used.

Data

Positive second. ion spectra;
measured at rectangular grid pos.,
$\Delta x = 15$ µm, $\Delta y = 10 - 30$ µm.
Mass range 12 – 72 Da
m = 665 variables (mass bins)

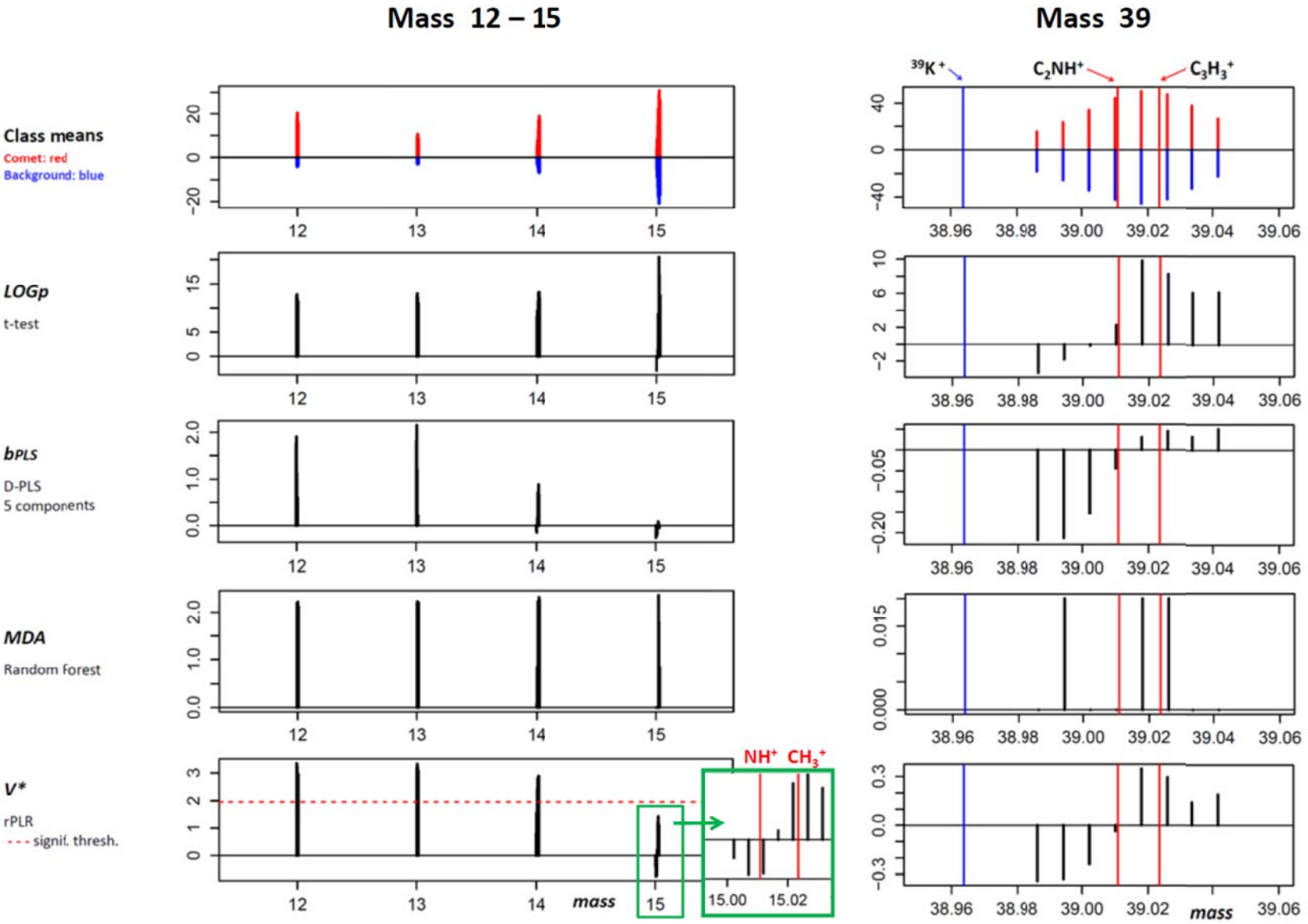
Spectra on particle: class 1
Particle *Sai*, target *3D1*
Size ca 80 µm
Collected 2016-04-13/14
n_1 = 29 spectra

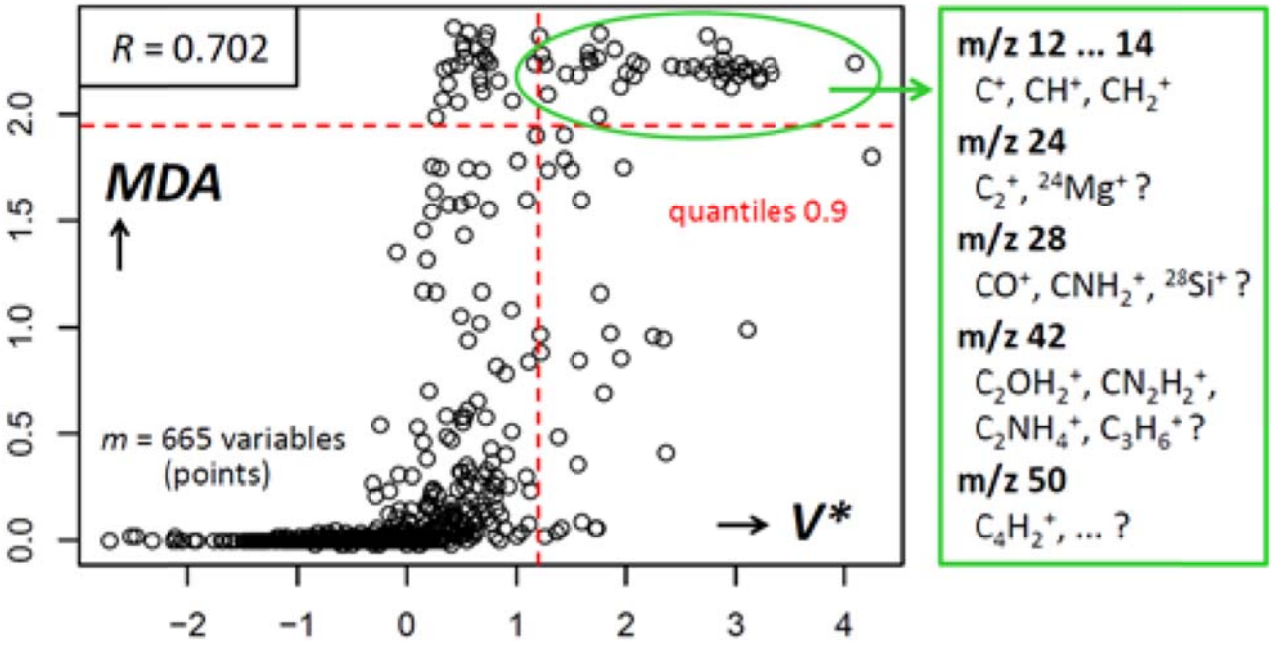


Spectra backgr: class 2
Area *Nick*, target *3D0*
Collected 2016-03-25/26
n_2 = 59 spectra

Selected Results

Variable importance measure versus ion mass





☞ Probably **from the cometary material**:

{ } not separable by mass; coded as **obvious** or as *guess*

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^+*, *$C_3H_4^+$*}; **$C_3H_3^+$**; ***C_4^+***; …

☞ Probably **from background**:

Saturated or lowly unsaturated CH ions:

e. g., $C_3H_{5\text{-}9}^+$, $C_4H_{7\text{-}9}^+$, $C_5H_{7\text{-}12}^+$

Summary

- Cometary particle surfaces contain CH(NO?) compounds.
- No distinct organic substance classes are evident from the data; a complex mixture of unsaturated organic compounds may be present.
- The results are consistent with the previously claimed presence of high molecular weight structures [8].

  - ❑ The applied methods for characterizing the variable importance are complementary.
  - ❑ Results from the univariate t-test are directly interpretable in terms of ion formulae. Results from the multivariate methods (D-PLS, RF, rPLR) characterize the variable importance within the potential influence of the other variables.

References

[1] http://www.esa.int/spaceinimages/Missions/Rosetta/(class)/image

[2] Schulz R., et al. (eds.): ROSETTA: ESA's mission to the origin of the solar system, Springer, New York (2009)

[3] Kissel J., et al.: Space Sci. Rev., **128**, 823 (2007)

[4] Langevin Y., et al.: Icarus, **271,** 76 (2016)

[5] Hilchenbach M., et al.: The Astrophysical Journal Letters, **816,** L32 (2016)

[6] Varmuza K., et al.: Chemom. Intell. Lab. Syst., **138**, 64 (2014)

[7] Walach J., et al.: submitted (2017)

[8] Fray N., et al.: Nature, **528**, 72 (2016)

This work is supported by the Austrian Science Fund (FWF), project P 26871 - N20.

Acknowledgments. COSIMA 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, 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.

J. W. thanks for support by the Austrian Science Fund (FWF) and the Czech Science Fund (GACR), project I 1910-N26.

Abstract

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

K. Varmuza[1*], P. Filzmoser[1], I. Hoffmann[1], J. Walach[1],
H. Cottin[2], N. Fray[2], C. Briois[3], J. Silén[4], O. Stenzel[5], J. Kissel[5], M. Hilchenbach[5]

[1] Vienna University of Technology, Vienna, Austria
[2] Université Paris Est Créteil et Université Paris Diderot, Créteil, France; [3] Université d´Orléans, Orléans, France
[4] Finnish Meteorological Institute, Helsinki, Finland; [5] Max-Planck-Institute for Solar System Research, Göttingen, Germany

* e-mail: kurt.varmuza@tuwien.ac.at; www.lcm.tuwien.ac.at/comecs/

The instrument COSIMA [1] on-board of the ESA mission Rosetta collected tens of thousand cometary dust particles at distances between 10 and some 100 km from the comet 67P/Churyumov-Gerasimenko. The particles were imaged [2], and the composition of the particle surfaces analysed by secondary ion mass spectrometry (SIMS) with a time-of-flight (TOF) mass analyser [3]. An essential aim of the project was obtaining information about presumable organic compounds on the comet.

Selected sets of mass spectra - either measured on the background or on cometary particles - have been transformed to data matrices suitable for binary multivariate classification. The variables used are ion counts in mass intervals (TOF time bins) relevant for ions containing C, H, and potentially N and O. Typically, the data sets contain 300 - 600 variables for the mass range 12 - 70 Da, and 30 - 500 objects (spectra) per class.

The instrumental mass resolution of about 1400 (at mass 100) allows a separation of elemental ions from H-rich organic ions; e. g., of $^{56}Fe^{+}$ (mass 55.935) from the possible nine CHNO-ions (giving overlapping peaks) in the neighbouring mass range 55.996 - 56.063. The ions from organic species around nominal mass 56 are distributed over about 20 TOF time bins (variables), and a deconvolution of the peaks is not feasible with the available data. The 20 variables can be considered as linear combinations of the signals from up to nine different CHNO ions.

The search for signals from any CHNO compound has been performed by applying methods that estimate the significance of variables for discriminating spectra from the background and from the particle surfaces. This task is similar to the search of biomarkers in high-dimensional data. Several methods have been tested: (1) comparison of the class means (of single variables) by the t-test or similar methods; (2) standardized regression coefficients of a discriminant variable obtained by D-PLS; (3) mean decreasing accuracies obtained by the random forest method; (4) robust pair-wise log-ratios [4]. The results are consistent with the previously claimed presence of high-molecular, carbon-rich material [5], and provide additional information about presumable carbon containing substances in cometary material.

.

[1] J. Kissel, et al., *Space Science Reviews,* **128** (2007**)** 823-867.
[2] Y. Langevin, et al., *Icarus*, **271** (2016) 76–97.
[3] M. Hilchenbach, et al., *The Astrophysical Journal Letters*, **816** (2016) L32.
[4] J. Walach, et. al., (2017) submitted.
[5] N. Fray, et al., *Nature,* **528** (2016) 72-74.