Automatic Machine Learning Classification Applied to Dawn/VIR data in view of MERTIS/BepiColombo M. D'Amore[1*], R. Le Scaon[2], J.Helbert[1], A. Maturilli[1], E. Palomba[3,4], A. Longobardo[3], H. Hiesinger[5], [1]German Aerospace Center DLR Berlin, Berlin (mario.damore@dlr.de), Germany, [2]Ecole Polytechnique, Université Paris-Saclay,Paris, France [3]IAPS-INAF, Rome, Italy , [4]ASDC-ASI, Rome, Italy , [5]Westfälische Wilhelms-Universität Münster, Germany.

Introduction: Remote sensing spectroscopy is one of the most commonly used technique in planetary science and for recent instruments producing huge amount of data, classic methods could fails to unlock the full scientific potential buried in these measurements. We explored several Machine Learning techniques: A multi-step clustering method is developed, using an image segmentation method, a stream algorithm, and hierarchical clustering.

The MErcury Radiometer and Thermal infrared Imaging Spectrometer (MERTIS) is part of the payload of the Mercury Planetary Orbiter spacecraft of the ESA-JAXA BepiColombo mission [1]. MERTIS's scientific goals are to spectrally identify rock-forming minerals, to map the surface composition, and to study surface temperature variations on Mercury. To cope with the stream of data that will be delivered by MERTIS, we developed an algorithm that could aggregate new data as they are acquired during the mission. This give the scientist a guide for the most interesting features on Mercury without being lost in high-volume dataset.

The NASA mission DAWN carries a suites of instruments aimed at understanding the two most massive objects in the main asteroid belt: Vesta and Ceres. DAWN has already successfully completed the exploration of Vesta in September 2012 and it is now in the extended mission phase around Ceres. The DAWN/VESTA VIR data are a testbed for the algorithm developed for MERTIS. The algorithm identified the olivine outcrops around two craters on Vesta's surface described in [2][3]. We furthermore mimic the data acquisition process as if the mission were dumping the data live with a data stream cluster algorithm, analyzing one data-cube and sequentially add the remaining data.

The algorithm provides insightful information on the novelty and classes in the data as they are collected. This will enhance MERTIS targeting and maximize its scientific return during BepiColombo mission at Mercury.

Method: In this work we analyzed photometrically corrected spectral parameter from DAWN/VIR spectral cubes, as defined in [4]. The parameter used are: infrared albedo, pyroxenes Band I (BI) and Band II (BII) band center positions, the fatality, forsterite, high-calcium pyroxene abundance indexes and the BII Half width at half maximum. Fig.1 shows the distribution maps of the fayalite index.

The analysis consists of three steps: 1. local homogenization, then 2. data stream clustering and 3. hierarchical clustering, all steps depending on previous steps results.

1. Mean-shift, a local homogenization technique was introduced in [5] as an image segmentation algorithm. In our work is applied as an iterative procedure based on kernel density estimation. Compared to the widely used k-means shows several advantages for our application. Applying both the techniques, Mean-Shift shows a higher stability in the number of cluster produced and segregates "noisy" and data pixel more efficiently.

2. After this step we apply a data stream cluster algorithm, the DenStream , introduced by [6]. The goal of this algorithm is to cluster a data stream, with no assumption over the number of clusters, or their shape, while being able to handle outliers. Our points are the cluster centers outputted by Mean-shift on each data

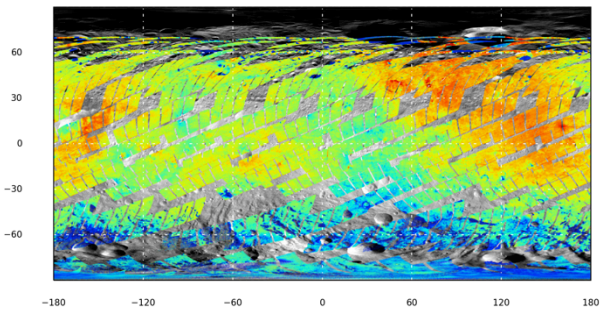

Fig.1 Fayalite Index distribution on the surface. [min=violet / max=red].

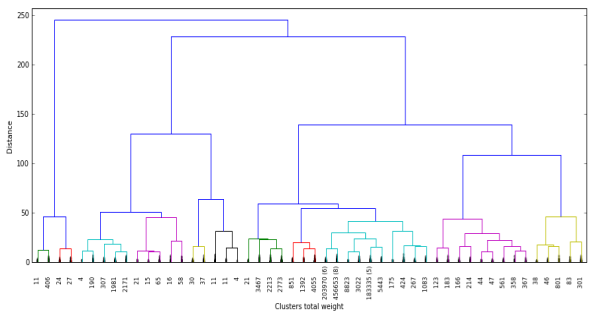

Fig.2 Last classification step: Hierarchical clustering. Dendrogram representing the last 50 clusters.

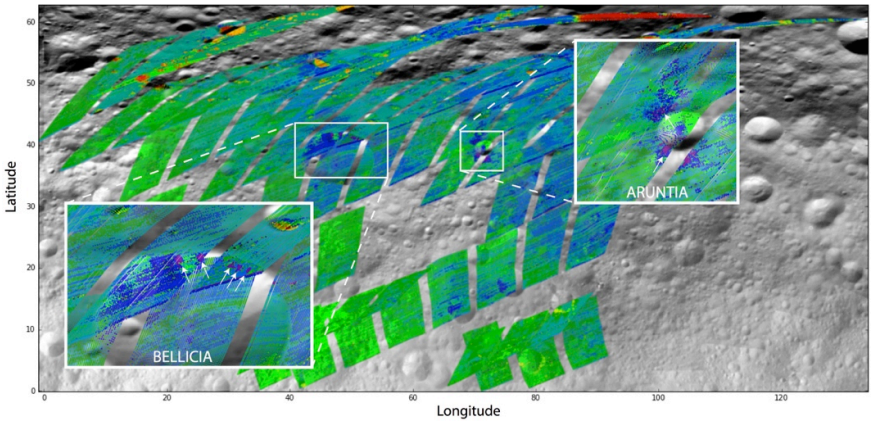

Fig.3 Map of classes after all the classification steps. Colors represent solely different classes.

cube, and are weighted by the number of measurements they represent. We classify the cluster from each datacube sequentially, assigning them to a big group of inlier cluster or outlier cluster, based on the cluster radius and total weight. This classification will be used during the active mission, as MERTIS data will be collected, to identify novelty in the datastream, either from new physical condition, instrument misbehavior or miscalibration, or residual of uncorrected photometrical effects. Abrupt increase in cluster numbers clearly mark points to investigate further, because indicates the input in the stream on unseen features in the data (i.e. novelty discovery, error in the data).

3. After the stream algorithm, a supplementary step can be necessary to get an overview of the distribution of clusters, as well as a simplified map of the surface. We chose to apply a hierarchical clustering analysis, using Ward's method to compute the linkage [7]. The Fig.2 shows a dendrogram representing the last 50 clusters. A vertical line corresponds to a cluster, and a horizontal line corresponds to a merging. The distance between two merged clusters is given by the height of the horizontal line. We notice that the core clusters from the stream algorithm are merged quite early, and are only shared between 3 of the 50 final clusters. The first 19 clusters starting from the left are likely to be measurement errors, as their distance to the core clusters is high. Mean-shift coupled with the stream algorithm reduces the data volume by a factor 500 in a reasonable time, and separates the most recurrent types of spectra from the more original ones, which helps spotting interesting anomalies on the surface or errors present in the data. Applying Hierarchical Clustering in the end allows a clearer understanding of the repartition of the data set.

Results: In order to evaluate objectively the quality of the clustering, we tried to reproduce the results of [2], who found outcrops of olivine around the two vestan craters Bellicia and Aruntia. The map in Fig.3 shows the datacubes colored as function of their cluster, after the Mean-shift and Stream algorithm steps. First, we notice the dark blue strips at the bottom of certain images. They are most likely caused by measurement errors that were clustered separately. The close-ups of Bellicia and Aruntia show magenta-colored clusters on the border of the craters. They spatially match the olivine-rich locations in [2]. We conclude that they mark the same compositional region on the surface, which in confirms the efficiency of the clustering method. In addition, by isolating those magenta clusters, it has been possible to identify other possible locations for olivine that will be studied in future works.

Conclusions: The workflow depicted here shows to be able 1. to reduce considerably the amount of data the scientist will deal with 2. to separate different features present in the data for easy inspection 3. to handle stream data, as expected in a live mission, moving novelty discovery before the whole dataset is acquired (typically this take severe years). The MERTIS will take advantage of this work automating the bulk work of data separation and giving more room for deep scientific analysis of the more interesting features on Mercury.

References: [1] H. Hiesinger et al (2010) PSS, 58, 1–2, 144–165 [2] Ammanito E. et al. (2013) , Nature, 504.7478, 122–125. [3] Ruesch O. et al. (2014), J. Geophys. Res. Planets, 119, 2078–2108. [4] Longobardo, A. et al (2014), Icarus, 240, 20–35. [5] Comaniciui, D. and Meer, P. (2002) PAMI , IEEE 24.5, 603–619. [6] Feng Cao et al. (2006) SDM, 6, 328–339. [7] Ward, J. H., Jr. (1963), JASA, 58, 236–244.