The impact of compositional changes on random forest predictions: applica-tion to chemcam libs data from gale crater, mars

K. Rammelkamp, O. Gasnault, C. C. Bedford, E. Dehouck, S. Schroder

THE IMPACT OF COMPOSITIONAL CHANGES ON RANDOM FOREST PREDICTIONS: APPLICATION TO CHEMCAM LIBS DATA FROM GALE CRATER, MARS K. Rammelkamp[1], O. Gasnault[2], C. C. Bedford[3,4], E. Dehouck[5], S. Schröder[1]; [1]German Aerospace Center (DLR), Institute of Optical Sensor Systems, Berlin, Germany; [2]Institut de Recherches en Astrophysique et Planétologie, Toulouse, France; [3]LPI, Universities Space Research Association, Houston, USA; [4]Astromaterials Research and Exploration Science, NASA Johnson Space Center, Houston, USA; [5]Université de Lyon, UCBL, ENSL, UJM, CNRS, LGL-TPE, Villeurbanne, France; kristin.rammelkamp@dlr.de

Introduction: In 2012, NASA's Mars Science Laboratory started its journey through Gale crater, Mars. One instrument of the payload is ChemCam which uses the LIBS (laser-induced breakdown spectroscopy) technique to analyze the chemical composition of the martian rocks and soils [1,2]. ChemCam is used nearly every sol and has collected a large dataset with more than 930 000 single shot spectra [3]. Such a dataset is an ideal candidate to be explored with statistical methods which can support the detection of changes in composition, alteration features or changes in sediment provenance. Unsupervised clustering is a common method which was used in several ChemCam analyses [4-6]. In one study, 6 clusters with dominant compositions were identified in the whole ChemCam dataset collected until sol 2756 [7]. The results were used as training data for a random forest (RF) model which reached high prediction accuracies and can be used to predict cluster memberships of new observations. But what happens to these predictions when the composition of new targets changes and is not covered anymore by the original 6 clusters? This question motivated the present study, in which we compare the RF predictions with results from unsupervised clustering of data from sol 3105-3580 (1664 observations) covering mainly the clay-sulfate transition [8].

Clustering Routine: For the new clustering, a similar approach as in [7] was employed: We did 100 runs of clustering where in each run, 1164 of the 1664 spectra were randomly selected. Non-negative matrix factorization (NMF) was applied to reduce the dimensionality prior to the k-means clustering. In order to identify the best number of NMF factors and clusters, the procedure was done for 3-8 NMF components and 3-8 numbers of clusters. The results were evaluated based on the silhouette score which is a measure for clustering quality [9]. The best mean silhouette scores among the runs were found for 5 NMF factors and 4 clusters. Closer inspection of this configuration reveals a strong consistency of the cluster centroids positions in NMF space and the cluster sizes among the 100 repetitions. Therefore, we decided for 5 NMF factors and 4 clusters. For the final cluster assignment, only those observations were kept which were always assigned to the same cluster among the runs. In the following, the results of this clustering will be denoted with *cluster x new*.

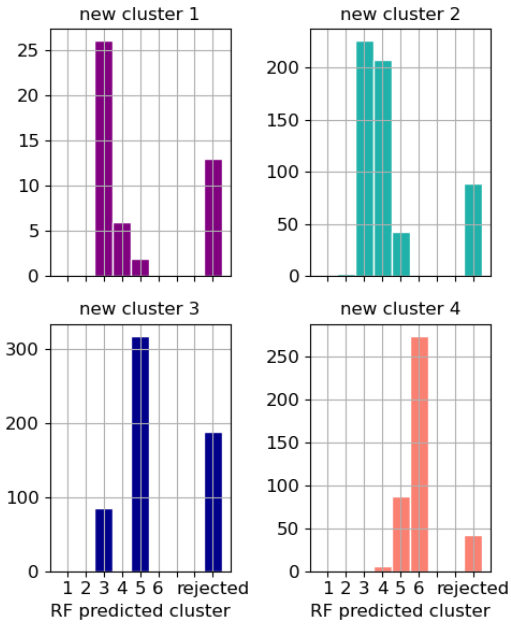

Figure 1: Histograms for each new cluster counting to which of the previous classes the RF predicted them.

Based on the NMF factors and comparison to the major oxide compositions [10], the dominant compositions can be broadly described as:

- cluster 1 new: High Na_2O and Cl (47 obs. points)
- cluster 2 new: High FeO_T and enrichments of elements typical for felsic rocks (SiO_2, Al_2O_3 and alkalis) (568 obs. points)
- cluster 3 new: MgO enriched (591 obs. points)
- cluster 4 new: CaO enriched (409 obs. points)

Comparison to RF Predictions: We used the RF model described in [7] to predict cluster memberships on the same dataset. In order to compare these predictions to the new clustering results, it was counted to which of the RF predicted clusters the new cluster members were assigned to. The results can be found in Figure 1 as histograms for each of the new clusters. Here, the term rejected means that the probability for the prediction by the RF was smaller than 0.7 and therefore rejected. Correspondences between new and RF predicted clusters can be observed: Predominantly, the new cluster 1 was predicted as RF cluster 3, new cluster 3 as RF cluster 5 and new cluster 4 as RF cluster 6. The new cluster 2 observations were classified as cluster 3 and 4 with approximately equal frequency by the RF.

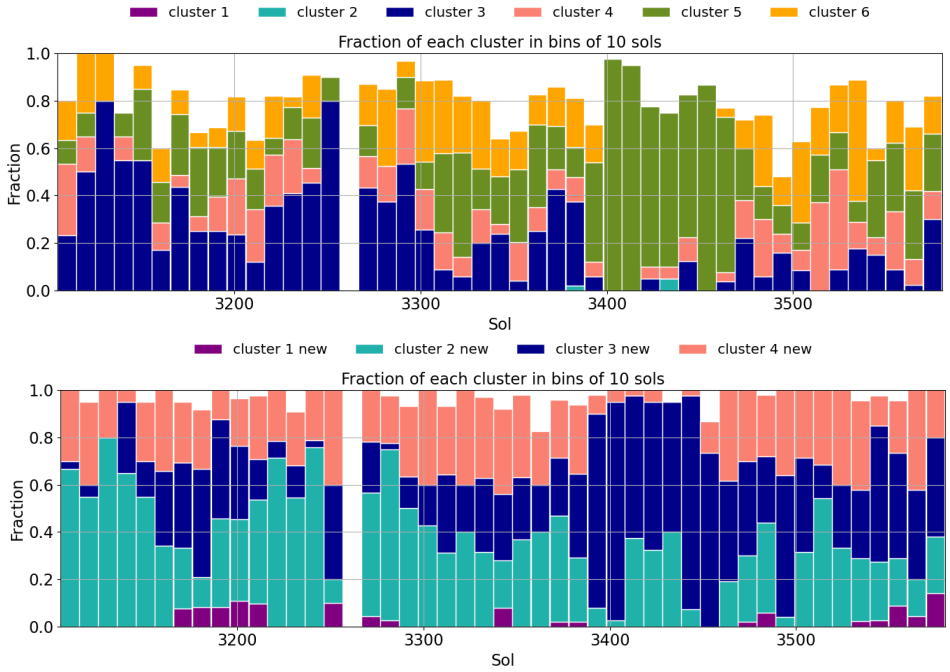

Figure 2: Both plots show proportion of each cluster observations in a constant bin of 10 sols. Rejected targets were counted, too. The upper figure shows the RF predictions, the lower one the results of the new clustering of the same interval of the mission.

Both new cluster 4 and RF cluster 6 can be described as high CaO clusters, thus, the correspondence between them is consistent. However, the new cluster 4 is on average lower in CaO (16.2 ± 5.8 wt %) than the high CaO cluster identified in [7] (23.1 ± 8.6 wt %). This shift is in agreement with the observation that Ca-sulfate is mostly present in smooth bedrock in the clay-sulfate transition [9] and in most cases not as pure veins.

Figure 2 shows for both the new clustering and the RF predictions the fraction of each cluster in bins of constant 10 sol width. Since the rejected observations are counted, too, it becomes apparent that in many bins of the RF predictions (upper plot), the proportion of rejections is 20 % or higher. This could already be an indicator that the new compositions do not always match the classes used for the training of the RF. In contrast, there are far fewer rejected observations with the new clusters, showing that they describe the dataset well.

In the time period of sol $\approx$ 3390-3470, in which Curiosity explored the younger Stimson formation on top of the Greenheugh pediment, the RF predicted cluster 5 and the new cluster 3 are dominant, respectively. Already during previous investigations of the Stimson formation this cluster 5 characterized as low SiO_2 was mainly observed [7] and also the correspondence between the two clusters is consistent. However, a small shift in composition between them can be observed: the new cluster 3 is somewhat lower in CaO and higher in MgO.

Another major difference between the two classifications is that no counterpart to the previous high SiO_2 cluster 1 was found in the new clustering. This was expected as the observations were strongly localized in the previous clustering study and also the RF did not predict any of this kind of observation. Also clear felsic compositions, cluster 2 in [7], were not predicted by the RF except for two observation points. Instead, a new cluster 1 with a clear halite contribution (high Na_2O and Cl detection) was identified.

Conclusions: The predictions of the RF model are still reasonable in most cases even though the composition in the clay-sulfate transition somewhat changed in comparison to previous formations most likely due to increased contributions of sulfates [9]. Observing the prediction probabilities can partly prevent strongly divergent predictions from being made by the RF. Nevertheless, in order to describe the compositions as accurately as possible, regular checks must be made, e.g. by monitoring the number of rejected observations and, if necessary, new models should be trained with extended as complete as possible training data.

References: [1] Maurice et al. (2012), *SSR*, 170; [2] Wiens et al. (2012), *SSR*, 170; [3] Gasnault et al. (2023), *this meeting*; [4] Gasnault et al. (2013), *44th LPSC*, #1994; [5] Gasnault et al. (2019), *9th Mars Conf.*, #6199; [6] Bedford et al. (2020), *Icarus*, 341; [7] Rammelkamp et al. (2021), *ESS*; [8] Rapin et al. (2023), *this meeting*; [9] Rousseeuw (1987), Journ. of Comp. and Appl. Math.; [10] Clegg et al., (2017), *SAP B*, 129, 64