OPTIMIZATION OF CLUSTERING ANALYSES FOR CLASSIFICATION OF CHEMCAM DATA FROM GALE CRATER, MARS K. Rammelkamp[1*], O. Gasnault[1], O. Forni[1], C. C. Bedford[2,3], E. Dehouck[4], J. Lasue[1], A. Cousin[1], S. Maurice[1], R. C. Wiens[5]; [1]Institut de Recherches en Astrophysique et Planétologie, Toulouse, France; [2]Lunar and Planetary Institute, Universitites Space Research Association, Houston, USA; [3]Astromaterials Research and Exploration Science, NASA Johnson Space Center, Houston, USA; [4]LGL-TPE, Lyon, France; [5]LANL, Los Alamos, USA; *kristin.rammelkamp@irap.omp.eu

Introduction: The ChemCam instrument on board the Mars science laboratory (MSL), which is the first extraterrestrial LIBS (laser-induced breakdown spectroscopy) instrument, has been successfully analyzing the martian surface in Gale crater since landing in 2012 [1-3]. LIBS is a rapid multi-elemental analysis technique and ChemCam is used nearly every sol which allows to track compositional variations on a small scale. Due to its regular use, ChemCam has collected a large dataset with more than 800 000 single shot LIBS spectra. Such a large dataset is suitable to use machine learning techniques for the identification of targets with similar compositions [4]. However, the training of those models is challenging due to the lack of training data from Mars. For particular ChemCam datasets, unsupervised techniques like hierarchical clustering were applied to support data interpretations and prove to provide conclusive classifications [5-7]. Here, the approach relies on the repeated application of k-means clustering on randomly selected sub-datasets of the whole ChemCam dataset. The objective of this study is to identify dominant compositions observed with ChemCam in Gale crater.

Method:

The dataset is limited to spectra measured at distances less than 3.5 m from the start of the mission until sol 2756 and contains 18719 spectra which are averages of usually 25 single shot spectra. In 100 runs, sub-datasets with 4896 spectra were randomly selected and on each sub-dataset the feature extraction method non-negative matrix factorization (NMF) was applied. We observed that six NMF factors were sufficient to describe the data. The 100 runs were repeated and on each sub-dataset k-means clustering with six clusters was applied to the NMF scores. Cluster assignments in each run were made based on mean NMF scores within the clusters. The comparison of the obtained clusters from each run revealed a strong consistency of cluster sizes (Fig. 1). As each observation was selected multiple times within the 100 runs (on average 25 times), only those observations were kept which were always assigned to the same cluster. This is the case for ≈ 92 % of the data confirming the stability of the approach. Nevertheless, in order to strengthen the distinction of the clusters from one another, we also evaluated cluster quality criteria like the silhouette score. Based on this criterion further observations were excluded from the analysis and interpretation of the six clusters. The final distribution of cluster members and out-sorted observations is shown in Fig. 2 where the different sizes of the clusters are observable.

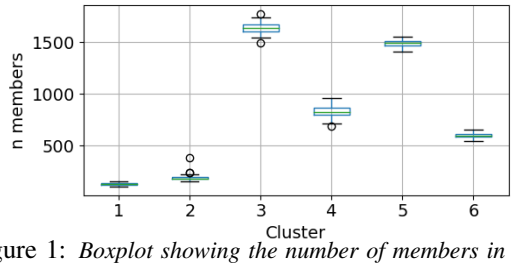

Figure 1: *Boxplot showing the number of members in each cluster for the 100 runs on sub-datasets. The cluster sizes are consistent among the repetitions, cluster 1 and cluster 2 are always rather small while cluster 3 and cluster 5 are always the largest clusters.*

Results: One of the results is that consistent clusters representing dominant chemical compositions could be identified in the ChemCam dataset. The frequency of cluster detections over the time of the mission are shown in Fig. 3 along with the elevation of the rover. Such a representation includes, besides the actual distribution of chemical compositions in Gale crater, any possible bias of target selection. In the following, each cluster will be briefly presented and discussed based on interpretations derived from the score values on the NMF factors, the observation frequency and also on comparisons with the major oxide compositions (MOC) [8].

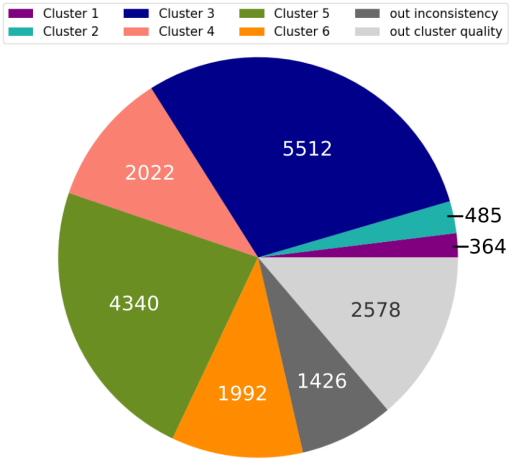

Figure 2: *Distributions of ChemCam spectra among the clusters. The greyish parts indicate observations that were not assigned to one of the six identified clusters either because they were not always assigned to the same cluster among the repetitions (inconsistency) or because of cluster quality criteria.*

Cluster 1: The smallest cluster consists of high silica targets (SiO_2 > 70 wt %) which were mostly detected at two locations between sol 1000 and 1500: Bridger Basin and Marias Pass. This is in agreement with the description of local diagenetic silica enrichment associated with late state groundwater activities in [9]. Cluster 1 observations at other locations, for example, after sol 2000 were made on a strongly altered target named *Askival* which has an igneous cumulate texture [10]. The Si-enrichment process of this target is most likely different from those observed at Bridger Basin and Marias Pass, even though they have the high silica content in common.

Cluster 2: This is the second smallest cluster whose members have their highest score values on a NMF component showing strong features of Al_2O_3, Na_2O and K_2O which indicates that the cluster is dominated by felsic compositions. These targets were detected mostly early in the mission and are associated with the crater rim and regional crust [11, 12].

Cluster 3: This is the largest cluster whose members were not regularly measured before ≈ sol 1400 when the rover reached the Karasburg member of the Murray formation. This cluster is interpreted to be largely representative of mudstones in the Murray formation.

Cluster 4: Due to high score values on a NMF factor showing strong Fe emission lines, members are expected to be enriched in FeO_T. Observations of this cluster were made rather constantly during the mission with some areas having an enhanced frequency of cluster 4 observations. For example, the peak of cluster 4 detections in the beginning of the mission, is in agreement with high FeO_T targets at Rocknest [13]. Furthermore, cluster 4 observations are more frequent during the time the rover explored the Vera Rubin Ridge (VRR) where local small FeO_T enrichments were targeted with ChemCam [14, 15].

Cluster 5: Like cluster 4, members of cluster 5 were rather constantly detected over the time of the mission. According to RMI image analysis and also from the MOCs we identified soil targets and sandstones in this cluster. Aeolis Palus and Bagnold Dunes soils [16] as well as observations from the Stimson formation, which consists of cross-bedded fine-grained sandstones and unconformably overlays the Murray formation [7], belong to this cluster. This becomes apparent, for example, during sols 2694-2733 when the rover investigated the Greenheugh Pediment which is part of the Stimson formation [17] by means of strongly reduced cluster 3 and enhanced cluster 5 observations during this time.

Cluster 6: Members in this cluster have a strong Ca component and can be associated with Ca-sulphate occurences as veins, cements and mixtures with the bedrock. The observation of Ca-sulphates started already in the beginning of the mission [18] but increased when the rover started to climb Mt. Sharp. Even though the majority of this cluster is associated with Ca-sulphates other high Ca targets like fluorites and apatites also belong to this cluster.

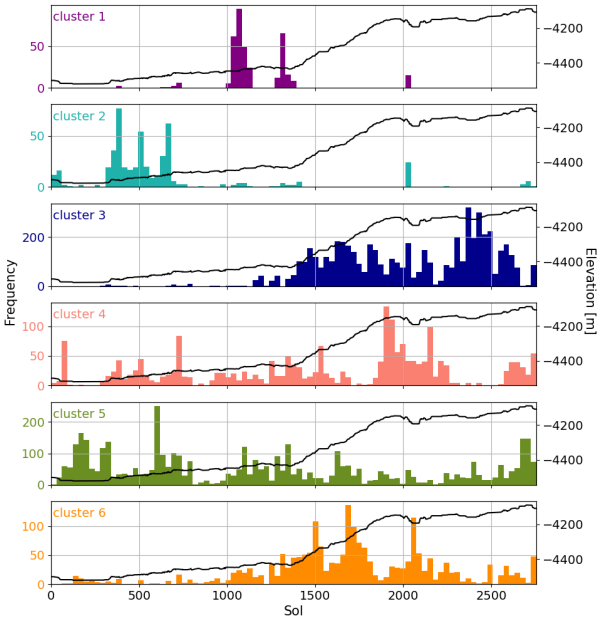

Figure 3: *Histograms of cluster observations over the time of the mission until sol 2756. In each plot, the binsize is ≈ 30 sols and the black curve indicates the elevation of the rover.*

Outlook: The six clusters represent dominant compositions observed in Gale crater with the ChemCam instrument, however, the whole chemical variability of martian compositions is not covered. Therefore, more detailed investigations of the clusters are currently made in order to sub-divide them into smaller sub-clusters. Preliminary results are encouraging for sub-clustering as well as mixing trends within the clusters. With this clustering study, machine learning classification models will be trained and validated which can support a rapid classification of new targets measured with ChemCam.

References: [1] Maurice et al. (2012), *SSR*, 170; [2] Wiens et al. (2012), *SSR*, 170; [3] Maurice et al. (2016), *JAAS*, 4; [4] Forni et al. (2019), *50th LPSC*, #1404; [5] Gasnault et al. (2013), *44th LPSC, #1994*; [6] Gasnault et al. (2019), *9th Mars Conf.*, #6199; [7] Bedford et al. (2020), *Icarus*, 341; [8] Clegg et al. (2017), SAP B, 129, 64; [9] Frydenvang et al. (2017), *GRL*, 44; [10] Bridges et al. (2019) 50th LPSC, #2345; [11] Sautter et al. (2015), *Nature Geosc.*, 8; [12] Cousin et al. (2017), *Icarus*, 288; [13] Blaney et al. (2014), *JGR: Planets*, 119; [14] L'Haridon et al. (2020), *JGR: Planets*, 125; [15] David et al. (2020), *JGR: Planets*, 125; [16] Cousin et al. (2017), *JGR: Planets*, 122; [17] Bedford et al. (2021), *this conference*; [18] Nachon et al. (2014), *JGR: Planets*, 119